Economic Theory (2007) 33: 145–167
DOI 10.1007/s00199-007-0212-0

SYMPOSIUM

Laurent Denant-Boemont · David Masclet · Charles N. Noussair

Punishment, counterpunishment and sanction enforcement in a social dilemma experiment

Received: 9 August 2005 / Revised: 8 January 2007 /
Published online: 31 January 2007

Abstract We present the results of an experiment that explores the sanctioning behavior of individuals who experience a social dilemma. In the game we study, players choose contribution levels to a public good and subsequently have multiple opportunities to reduce the earnings of the other members of the group. The treatments vary in terms of individuals' opportunities to (a) avenge sanctions that have been directed toward themselves, and (b) punish others' sanctioning behavior with respect to third parties. We find that individuals do avenge sanctions they have received, and this serves to decrease contribution levels. They also punish those who fail to sanction third parties, but the resulting increase in contributions is smaller than the decrease the avenging of sanctions induces. When there are five rounds of unrestricted sanctioning, contributions and welfare are significantly lower than when only one round of sanctioning opportunities exists, and welfare is lower than at a benchmark of zero cooperation.

We thank James Andreoni, participants in seminars at Emory University, the University of Wisconsin-Madison, the University of New South Wales, the University of Sydney, Deakin University, the 2004 North American Regional Meetings of the ESA in Tucson, Arizona, USA, the 2004 IMEBE Meetings in Cordoba, Spain, and the 2005 SAET meetings in Vigo, Spain, for constructive and helpful comments. We thank Elven Priour for programming and organization of the sessions. Instructions for the experiment are available from the authors.

C. N. Noussair (✉)
Department of Economics, Faculty of Economics and Business Administration,
Tilburg University, P. O. Box 90153, 5000 LE Tilburg, The Netherlands
E-mail: C.N.Noussair@uvt.nl
Tel.: +31-13-4662690

L. Denant-Boemont · D. Masclet
Department of Economics, Université Rennes 1, Rennes, France

D. Masclet
CIRANO, Montréal, Canada

Keywords Public goods · Sanction enforcement · Counter punishment · Information

JEL Classification Numbers C92 · D70 · H41

1 Introduction

One focus of experimental research on social dilemmas has been the search for and the identification of factors that promote cooperative behavior in settings in which individuals have incentives to behave opportunistically. The most widely used arena for this investigation is the voluntary contributions mechanism, an elegant construction that permits straightforward measurement of the extent of self-versus group-interested behavior. Interaction in the voluntary contributions mechanism proceeds according to the following rules. Each member of a group of individuals has an endowment, from which he may contribute any amount to a public good that returns a payoff to each individual. The level of this payoff ensures that at the social optimum, each individual contributes his entire endowment while, in contrast, each individual has a dominant strategy to contribute zero. The amount contributed can be interpreted as a measure of cooperative behavior. The main overall pattern observed in laboratory experiments is that initial contributions are substantial, but decline as the game is repeated and cooperation converges to a near-negligible level in the long run (Isaac et al. 1985; Andreoni 1988; Isaac and Walker 1998a; Ledyard 1995).

However, a number of modifications to the game that increase cooperation considerably, even in the long run, have been identified.[1] For example, endowing individuals with the ability to reduce the earnings of the least cooperative individuals in the group is highly effective in raising contribution levels (see for example Yamagishi (1986); Ostrom et al. (1992); Fehr and Gaechter (2000); Carpenter (2005b); Bochet et al. (2005); Masclet et al. (2003); Noussair and Tucker (2005)). These studies and others all find that individuals are willing to pay from their own earnings to reduce the earnings of free riders, and average contributions increase as a result of the existence of the sanctioning opportunity. It is thus clear that, at least under some circumstances, sanctioning mechanisms can represent an effective means of increasing cooperation among individuals and thus alleviate free-rider problems. This is the case even when the punishment is costly for sanctioners to administer,[2] and when the system does not rely on an external trigger mechanism for enforcement.

Immunity of sanctioners from reprisals is a characteristic of all of the studies listed in the last paragraph. Because there is only one opportunity to sanction in each period, and there is no means to track the identity of others from period to period, no player can identify individual punishment behavior in a manner that

[1] These include preplay communication (Isaac and Walker 1988b), creation of group identification in conjunction with post-play open discussion (Gaechter and Fehr 1999), and having each individual assign a rating to each other group member's contribution decisions (Masclet et al. 2003).

[2] See Falk et al. (2005) for a detailed analysis of the motivation behind the application of costly sanctions.

allows him to target an individual for reciprocation.[3] If such reprisals were possible, it might deter sanctioning, and thus dilute the effectiveness of the system in increasing contribution levels. Nikiforakis (2004) reports an experiment focused on this issue. He conducts an experiment in which there are two rounds of sanctions. Each individual becomes aware, after the first round of sanctions, of the punishment that each individual assigned to him. He then has the opportunity to sanction those, but only those, who sanctioned him. This creates a second round of sanctions, but only for the purpose of avenging sanctions received, which is termed as *counterpunishment*. Nikiforakis finds that the existence of the option to counterpunish nearly entirely offsets the increase in contributions the existence of the opportunity to punish creates.

On the other hand, all of the above studies preclude the use of punishment to attempt to refocus sanctions on low contributors, with the goal of increasing cooperation. There are two principal mechanisms for using sanctions in this manner. The first is to penalize those who fail to punish low contributors, and the second is to sanction those who punish high contributors. We will refer to these forms of second order sanctioning as "sanction enforcement".[4] Cinyabuguma et al. (2004, 2005) report an experiment in which individuals observe how much of other players' punishment assignments over the previous three periods (of play of a repeated game) are directed toward above-average, below-average, and average contributors. In every third period there is a second round of punishment, in which players can condition their second round of punishment on, but only on, this information. Agents cannot counterpunish those who assigned them sanctions because they do not observe the identity of the individuals to whom sanctions are directed. While those engaging in "perverse punishment",[5] the punishment of high contributors, are targeted most severely during the second opportunity to sanction, even those who sanctioned low contributors were punished more than those who did not. When a second opportunity to sanction is available, it is characterized by higher contributions and earnings compared to a setting, in which no second punishment opportunity exists. The data of Cinyabuguma et al. suggest that instead of increasing cooperation by allowing punishment of those who fail to sanction, the second stage of sanctions increases contributions and earnings by deterring punishment of high contributors.

The Nikiforakis and Cinyabuguma et al. studies find that counterpunishment reduces, and sanction enforcement increases, contributions. Our design, described in detail in Sect. 2, allows us to consider the following two issues that these ear-

[3] In some of the studies, in which group membership was fixed, agents could punish all other group members by contributing less or by randomly sanctioning other agents in subsequent rounds, but an individual could not be targeted for sanctions based on his prior sanctioning behavior.

[4] See Coleman (1990) for a detailed discussion of the use of sanctioning mechanisms to promote and sustain cooperation. On the issue of sanction enforcement, he writes "If the second order sanction is a positive one, reward to the sanctioner must be provided whenever the right action (sanctioning the initial offender) is taken; a negative sanction must be applied only when the wrong action is taken. If there develops a norm that one must sanction the violator of the initial norm, then the negative second order sanction for not applying the first order sanction must be applied only when that sanctioning norm is violated. This cost reduction to norm beneficiaries may give them an interest in establishing a sanctioning norm".

[5] The incidence of sanctioning of high contributors depends on the subject pool employed (Gaechter and Herrmann, 2005).

lier studies leave unaddressed. The first issue is which effect, counterpunishment or sanction enforcement, is greater in magnitude. This cannot be deduced from comparing previous studies, because they differ from each other in the timing and parametric structure of the interaction studied. We investigate this issue in two ways. The first method is to construct a treatment, in which all individual sanctioning and contribution behavior is observable to all players, and a second round of sanctions exists. This allows the net effect of the two forces on contributions to be measured. The second method is to isolate the effect of counterpunishment and of sanction enforcement by constructing two treatments, which are identical except that only one of the two behaviors can occur, and to compare the magnitudes of their effects on contributions. As described in Sect. 3, we find that the effect of counterpunishment is greater than that of sanction enforcement, and that the addition of a second round of unrestricted sanctioning reduces contributions and welfare. The second issue is the effect of multiple additional opportunities to sanction. We consider this issue by studying a game in which there are five rounds of sanctions in each period, the full history of sanctioning in earlier rounds is available, and any player may punish any other. We find that both contributions and welfare levels are significantly lower than when there is only one round of sanctions. Indeed, the level of welfare is lower than the minimum possible level in the absence of a sanctioning mechanism. Thus, the introduction of additional rounds of sanctioning, with all of the complex sanctioning strategies they enable, magnifies the effect of adding a second round of sanctions. The details of our experimental design are presented in Sect. 2. We present the results of the experiment in section three, and we offer a summary and some thoughts on the theoretical modeling of the patterns we observe in Sect. 4.

2 The experiment

2.1 Overview

There are five treatments in the experiment, all of which have a first and a second stage of interaction in common. In the first stage of the game, players simultaneously decide how much of their endowment to contribute to the public good. In the second stage, players are informed of the decisions that other members of their group have made and have the opportunity to punish them. The punishment reduces the earnings of both the sanctioning and the sanctioned parties. In the *Baseline* treatment, the two stages described above comprise all of the activity in the game. The Baseline treatment is a replication of Fehr and Gaechter (2000). In the other treatments, in contrast, there are additional stages, in which players are informed about some or all of the sanctioning activity in stage two or subsequent stages, and any individual may sanction any or all members of his group in a manner similar to stage two.

Three of the remaining treatments consist of exactly three stages, one contribution and two sanctioning stages, and differ from each other in terms of the information available to players about the identity of those administering sanctions in the second stage (the first sanctioning opportunity). In the *Full Information* treatment, players are informed about how much each individual sanctioned each other individual. In the *Revenge Only* treatment, each individual is informed of the source and the quantity of the sanctions directed toward him, but does not know

how much other members of the group were sanctioned and by whom. In the *No Revenge* treatment, no individual is informed about who sanctioned him personally and by how much. However, all individuals are informed of the source and the quantity of the sanctions directed toward each player other than himself. The information received at the end of the second stage may be used in the determination of punishment assignment strategies in the third stage.

The Baseline treatment allows punishment only in response to contribution decisions. The Revenge Only treatment allows counterpunishment as well as punishment for contribution behavior in the first stage. The No Revenge treatment allows sanction enforcement and punishment for low contributions, but precludes counterpunishment (although individuals may form beliefs about who sanctioned them based on who sanctioned others, and may attempt to counterpunish based on this information). The Full Information treatment allows sanction enforcement, counterpunishment, and punishment for low contributions. Therefore, the difference in contributions between the Baseline and the Revenge Only treatments, as well as the difference between the No Revenge and the Full Information treatments, measure the marginal effect of counterpunishment on contributions. The first and second comparisons are within an environment in which sanction enforcement is impossible and possible, respectively. We hypothesize, based on the results of Nikiforakis (2004), that the effect of the possibility of counterpunishment on contributions is negative.

The difference in contributions between the Baseline and the No Revenge treatments, as well as the difference between the Revenge Only and the Full Information treatments, is interpreted as the effect of the introduction of sanction enforcement. We hypothesize, based on the results of Cinyabuguma et al. that the effect of sanction enforcement on contributions is positive. The difference between the Full Information and the Baseline treatments measures the effect of removing all immunity from reprisals for sanctioning decisions, thereby incorporating the effects of both counterpunishment and sanction enforcement. The overall effect of these forces is ambiguous in sign. To summarize, our hypotheses concerning treatment differences are those specified in (1).

$$C(NR) > C(B), \quad C(NR) > C(FI), \quad C(FI) > C(RO), \quad C(B) > C(RO), \quad (1)$$

where $C(X)$ refers to the average amount contributed in treatment X, and NR, B, FI, and RO are abbreviations for the four treatments. The final treatment is the six-stage full Information treatment (6SFI). In this treatment, as in the other four, the first stage of the game consists of players simultaneously choosing how much of their endowment to contribute to a public good. In the second stage, players are informed of the decisions that other members of their group have made and have the opportunity to punish them, as in the Full Information treatment. Stages 3–6 are identical to stage 2. Players observe the sanctioning decisions of all members of the group in all earlier stages and then may sanction any other individuals. The 6SFI treatment allows punishment for complex motivations such as punishing counterpunishment or punishing a failure to enforce sanctions. The changes in contributions that the additional motives to punish that the extra stages introduce would appear to be ambiguous in sign, and thus we advance no hypotheses ex-ante about whether 6SFI leads to lower or higher contributions than the Baseline or Full Information treatments.

2.2 Procedures

The experiment consists of ten sessions, two sessions conducted under each of the five treatments. An average of twelve individuals participated in each session, for a total of 120 participants.[6] All sessions were conducted at the LABEX of the University of Rennes I, Rennes, France in 2004. The experiment was computerized and the scripts were programmed using the z-tree platform (Fischbacher, 1999). The subjects were undergraduate students from a variety of majors. Roughly one-third were economics students in the first 2 years of their studies at the University, and all but a small number of the remaining two-thirds were students in law, management, and medicine. No individual participated in more than one session.

In each session, there are 20 periods of interaction. Each period within a session proceeds under identical rules. The subjects participating in the session are assigned to groups of size four with fixed membership, in such a manner that they do not know the identities of the other members of their group. There were six groups, and thus six independent observations, for each treatment. At the end of each period, individuals remain in the same group. However, individuals' designated labels and the location of the display of their data on the computer screen are reassigned on a random basis in each period. For example, if a player is designated as player A in period t, he has exactly a one-fourth chance of being player A in period $t + 1$, as well as a one-fourth chance of being player B, C, *or* D in period $t + 1$.

The design and the parametric structure of the experiment draw heavily on those of Fehr and Gaechter (2000). At the beginning of each period in all treatments, each participant receives an endowment of 20 ECUs (experimental currency units, with 1 ECU = 2 Eurocents). He then must choose to allocate the endowment between a private account, which is his to keep, and a public account, which yields. 4 ECUs to each member of the group for each ECU allocated to the account by any group member. Following previous authors, we will refer the amount that the individual allocates to the group account as his *contribution*, because the more he allocates to the group account, the lower his own but the greater the group's total earnings. At the end of the first stage, each individual's provisional earnings are equal to

$$\pi_i^1 = (20 - c_i) + 0.4 \sum_{j=1}^{4} c_j,$$

where c_i is the contribution of player i.

After contribution choices have been made, they are revealed to all group members, and the game enters stage two. Each member of the group is informed of the total contribution of the group and the individual contribution of the three other group members to the public good, as well as her own provisional earnings after the first stage, π_i^1. In stage two, players have an opportunity to assign sanctions to each of the other members of their group. Sanctions take the form of an assignment of a number of punishment points in the range from 0 to 10 inclusive. A different number of points and thus a different sanction level may be assigned to each

[6] In eight of the ten sessions, there were exactly twelve participants, divided into three groups of four. In one of the Six Stage Full Information sessions there were eight participants, comprising two groups, and in the other session there were sixteen participants and thus four groups.

other player. Assigning points is also costly to the sanctioner. The cost function for punishment for player i is denoted as $k_i(p_i^{jm})$, where p_i^{jm} is the number of points that player i assigns to j in stage m.[7] The cost function for punishment points is defined for each pair of individuals so that $k_i\left(\sum_j p_i^{j2}\right) = \sum_j k_i\left(p_i^{j2}\right)$, The cost function is common to all individuals, so that $k_i\left(p_i^{j2}\right) = k_n\left(p_n^{j2}\right) = k\left(p_i^{j2}\right)$. Player i's provisional earnings after the second stage are given by:

$$\pi_i^2 = \pi_i^1\left[\max\{0,\, 1 - (1/10)\sum_{j\neq i} p_j^{i2})\}\right] - \sum_{j\neq i} k\left(p_i^{j2}\right). \tag{2}$$

Each point the agent receives reduces his earnings by 10% of stage 1 earnings, with a maximum reduction of 100% (receiving more than 10 points imposes no further reduction in earnings, but nonetheless is costly to sanctioners). The cost of punishment is then subtracted to calculate provisional earnings after stage two. The two stages described above comprise all of the activity in a period of the Baseline treatment.

In the four other treatments, there are subsequent stages of activity. The FI, NR, and RO treatments consist of three stages while 6SFI consists of six stages. The FI, NR, and RO treatments differ only in the information available to each individual after stage two. In the Full Information treatment, each player i is informed of the amount that each player sanctioned each other player. That is, he observes p_k^{j2}, for all j and k. In the No Revenge treatment, player i is informed only about how other individuals were sanctioned. That is, player i observes p_k^{j2}, for all k and for all $j \neq i$, but not for $j = i$. In the Revenge Only treatment, each player is only informed about his own sanctions received. In other words, individual i observes p_k^{i2} for all k, but does not observe p_k^{j2} for $j \neq i$. The cost of the sanctions to both the sanctioning and the sanctioned parties is identical to stage two in the three treatments. That is $k(p_i^{j3}) = k(p_i^{j2})$. In the three treatments, subject i observes the total number of points assigned to him in each of the two punishment stages.

After the appropriate sanctioning information is transmitted to participants, each member of the group has a second opportunity to assign punishment points to the other players. As in the earlier punishment stage, each point received reduces an individual's earnings by 10% of her first stage earnings. The final earnings for individual i in a period of the NR, FI, and RO treatments are equal to:

$$\pi_i^3 = \pi_i^1\left[\max\{0,\, 1 - (1/10)[\sum_{j\neq i} p_j^{i2} + \sum_{j\neq i} p_j^{i3}]\}\right] - \sum_{j\neq i} k\left(p_i^{j2}\right) - \sum_{j\neq i} k\left(p_i^{j3}\right) \tag{3}$$

In the 6SFI treatment, there are three more stages that follow stage three. These are identical to stages two and three of the FI treatment in that at the end of each stage

[7] The cost function $k_i(p_i^{jm})$ used in the experiment corresponded to a marginal cost of c ECU for $2c - 2 < p_i^{jm} \leq 2c$, and is identical to the function used in Fehr and Gaechter (2000) and in Masclet et al. (2003).

s, subject i observes all individuals' sanctioning behavior $p_j^{k,r}$ for all j, k and for all prior stages $r < s$. The payoff function at the end of stage 6 for individual i equals:

$$\pi_i^6 = \pi_i^1 \left[\max\{0, 1 - (1/10) \sum_{s=2}^{6} \sum_{j \neq i} p_j^{is}\} \right] - \sum_{s=2}^{6} \sum_{j \neq i} k \left(p_i^{js} \right) \qquad (4)$$

At the end of each period in all treatments, each participant's computer displays the contribution of each individual, the sanctioning information about others permitted under the treatment condition, the total quantity of punishment points the individual received in each stage, and his period and accumulated earnings for the session.

3 Results

This section has the following structure. Section 3.1 reports the data from the treatments with two sanctioning stages, and presents evidence for the first main result of the study. This is that the negative effect on contributions of permitting counter-punishment is greater in magnitude than the positive effect of permitting sanction enforcement. Section 3.2 considers individual sanctioning patterns in detail and Section 3.3 studies the subsequent response of recipients of sanctions. Section 3.4 presents the results from the 6SFI treatment, and details the evidence for the other principal result of the study. This is that the addition of multiple additional rounds of sanctions reduces contributions compared to a setting with only one stage of sanctioning, and also reduces welfare relative to a setting with no sanctions or with one stage of sanctions.

3.1 The effect of a second punishment stage on contribution and welfare levels

Figure 1 illustrates the time path of individual contributions by period, averaged across groups, in the five treatments. The period number is shown on the horizontal axis and the average contribution on the vertical axis. The maximum possible individual contribution, corresponding to the group optimum, is 20. The minimum possible contribution is 0. Figure 2 shows the corresponding time series of average individual earnings by treatment. The maximum possible individual earnings level, associated with all players contributing their entire endowment and no sanctioning, is 32, while individual earnings equal 20 if all contributions are zero. The average contribution for each group in each treatment is shown in Table 1, with the standard deviations given in parentheses.

Average contributions are highest in the No-Revenge treatment (16.17 per individual from a maximum possible of 20), followed in turn by the Baseline (15.49), the Full Information (10.59) and the Revenge Only (7.21) treatments, but there is considerable heterogeneity between groups in all of the treatments. The treatment differences are all consistent with the hypothesized orderings stated in (1). In the No-Revenge Treatment, 5 of 6 groups contribute more than 85% of their endowment over the 20 periods, while no group in the Full Information or Revenge Only treatments does so. On average, contributions in Revenge Only are less than half of the levels in the Baseline and the No Revenge treatments. Figure 1 suggests that

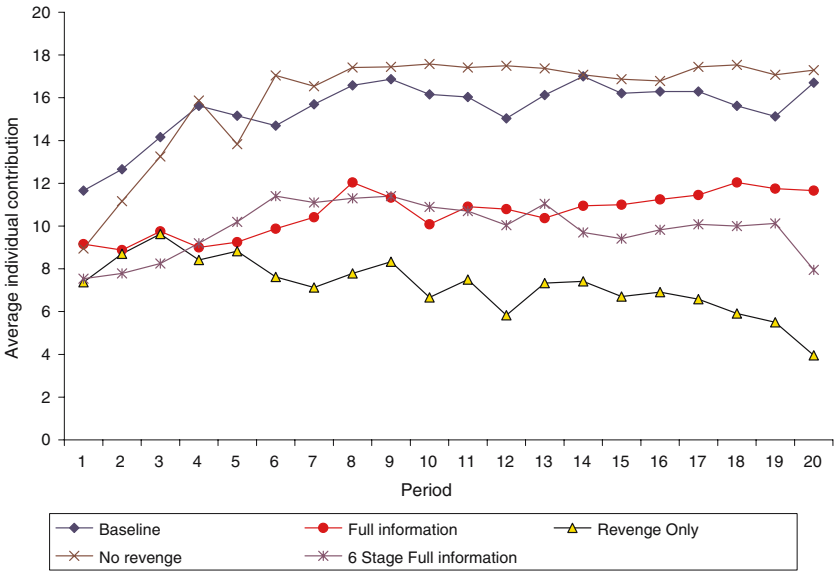

Fig. 1 Average individual contribution levels in each treatment

the average contribution level does not change appreciably as the game is repeated in any treatment, with the exception of a decline over time in Revenge Only, and an initial increase in the first few periods of Baseline and No Revenge.

A Mann–Whitney pairwise statistical test comparing contributions between treatments, maintaining the conservative assumption that each group's activity over the session is a unit of observation, yields the results shown in Table 2. The unit of observation is the average contribution of the group over the entire session (yielding six observations per treatment, one per group), and the null hypotheses are that the median group in each of the two treatments contributes an identical amount over a 20-period session. Identical inferences relative to the critical values for $p = 0.05$ are obtained if the data from the last five periods rather than those from entire sessions are compared.[8]

Introducing the possibility of counterpunishment has the effect of reducing contribution levels, whether or not sanction enforcement is possible. The difference in contributions between the Baseline and the Revenge Only treatments, as well as the difference between the No-Revenge and the Full Information treatments, is significant. Thus, we observe a similar effect as Nikiforakis (2004), in that counterpunishment reduces contributions, and we find that it generalizes to a setting in which sanction enforcement exists. Sanction enforcement has a positive, but not significant, effect in increasing contributions. This effect is similar to that observed by Cinyabuguma et al. (2004, 2005).The differences between the Full Information and the Revenge Only treatments and between the Baseline and the No-Revenge treatments are not significant. The (borderline, $p < 0.1$) significant difference between the Baseline and the Full Information treatments indicates that the effect of sanction enforcement is not strong enough to fully offset the decrease

[8] The tests used have relatively low power to reject null hypotheses of no difference between treatments compared to standard parametric methods.

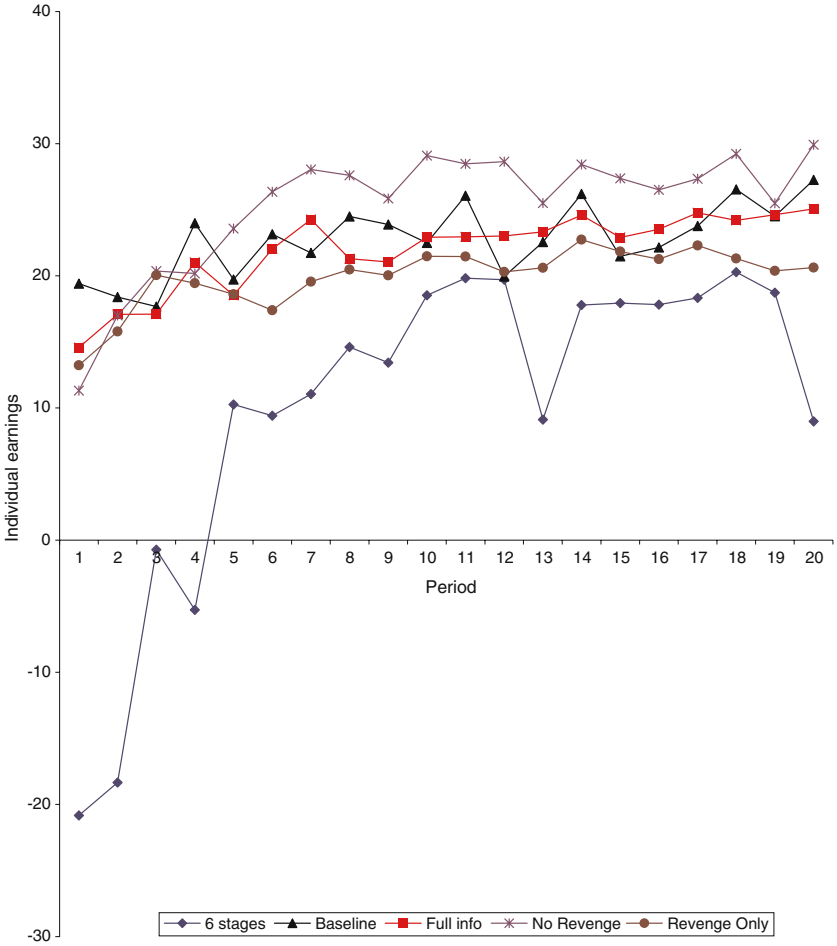

Fig. 2 Average individual earnings in each treatment, by period

in contributions from counterpunishment. Thus, allowing unrestricted reprisals for one round of sanctions reduces overall contributions, though the effect is only of borderline significance.

The introduction of the ability to engage in sanction enforcement, in conjunction with a prohibition on counterpunishment, is welfare-improving. The difference in total earnings between the Revenge Only and the No Revenge treatments is significant at the $p < 0.05$ level, according to a Mann–Whitney rank sum test, with the No Revenge treatment leading to the higher earnings. The same results are obtained when the last 5 periods are used in the test rather than all 20 periods. Average earnings per period in the Revenge Only treatment, 19.94 tokens, are comparable to those that would result if no individual made any contributions and there were no sanctions possible (20 tokens). The ability to enforce sanctions on its own does not increase welfare. The No-Revenge treatment does not generate significantly higher earnings than the Baseline treatment, and similarly, there is no

Table 1 Average individual contribution levels by group in each treatment (max = 20, min = 0)

	Baseline	Full information	Revenge only	No revenge	Six stage full information
Group 1	11.5	12.93	6.225	7.47	12.33
	(3.44)	(4.20)	(2.42)	(4.38)	(5.41)
Group 2	16.73	3.42	16.01	18.85	4.48
	(4.71)	(5.75)	(6.15)	(3.13)	(5.35)
Group 3	17.06	12.65	2.275	18.85	16.05
	(4.79)	(3.46)	(2.59)	(3.24)	(5.13)
Group 4	18.03	5.575	2.15	17.31	10.4
	(2.504)	(6.37)	(1.51)	(5.39)	(3.89)
Group 5	10.63	12.52	5.45	17.33	5.49
	(3.40)	(3.52)	(3.20)	(4.87)	(2.09)
Group 6	18.96	16.47	11.13	17.23	10.89
	(2.14)	(5.13)	(6.68)	(4.03)	(3.00)
Average	15.49	10.59	7.21	16.17	9.93
Std. dev.	(3.50)	(4.74)	(3.75)	(4.17)	(5.87)

Table 2 Results of Mann–Whitney rank sum tests of differences in contribution levels and earnings between treatments (level of confidence at which null hypothesis of no difference between treatments can be rejected, each session mean is a unit of observation)

Contributions	Full information	Revenge only	No revenge	6 Stage full information
Baseline	$p < 0.10$	$p < 0.01$	Not sig.	$p < 0.05$
Full Inf	–	Not sig.	$p < 0.01$	Not sig.
Revenge	–	–	$p < 0.005$	Not sig.
No revenge	–	–	–	$p < 0.05$
Earnings				
Baseline	Not sig.	$p < 0.1$	Not sig.	$p < 0.02$
Full Inf	–	Not sig.	Not sig.	$p < 0.05$
Revenge	–	–	$p < 0.05$	$p < 0.05$
No revenge	–	–	–	$p < 0.01$

significant difference between Revenge Only and Full Information. On the other hand, there is some evidence that the possibility of counterpunishment is welfare-reducing. Earnings under the Revenge Only treatment are lower than under the Baseline treatment and the effect is borderline significant. Average earnings are not significantly different in the Baseline and Full Information treatments.

The average quantity of sanctions one individual assigns to another in each stage of the four treatments is shown in Table 3. The No Revenge, Revenge Only and Full Information treatments have a similar overall total number of sanctions applied. This total is less than in the Baseline treatment, despite the fact that the Baseline treatment has one fewer round of sanctions.

3.2 Who sanctions whom?

When they make their decisions in stage two, agents have observed the contribution decisions of all other individuals, and can condition their sanctioning behavior on this information. The tendency to punish those who contribute less than the

Table 3 Average quantity of sanctions $\left(\sum_i \sum_k p_i^{kmt}/n\right)$ assigned by individuals to the rest of the group in each stage of a period

Treatment	Baseline	Full information	No revenge	Revenge only	Six stage full information
Average points assigned in stage 5					0.769
Average points assigned in stage 4	/	/	/	/	0.650
Average points assigned in stage 3	/	/	/	/	0.544
Average points assigned in stage 2	/	0.57	0.37	0.38	0.527
Average points assigned in stage 1	1.512	0.46	0.65	0.73	1.617
Average assigned over the stages	1.512	1.03	1.02	1.11	4.11

group average can be seen in Table 4. The left side of the table reports the results from the estimation of Eq. (5) for the data from stage two for all periods of the Full Information, No Revenge, and Revenge Only treatments. The right side of the table shows the data from period 1 of the three treatments. The regression is also conducted for the data from $t = 1$ alone to provide estimates in which there is no possibility for endogeneity, due to the dependence of the independent variables in period $t + 1$ on the dependent variables in period t, to influence inferences

$$p_i^{j2t} = \beta_0 + \beta_1 \bar{c}^t_{-i} + \beta_2 \max\left\{0, \bar{c}^t_{-j} - c^t_j\right\} + \beta_3 \max\left\{0, c^t_j - \bar{c}^t_{-j}\right\} + \beta_4 t. \tag{5}$$

The dependent variable p_i^{j2t} is the quantity of punishment points that player i assigns to player j in the second stage of period t, c^t_j is player j's contribution in period t, and $\bar{c}^t_{-j}$ is the average contribution of individuals *other than* j in period t.[9] A significantly positive coefficient on $\beta_2(\beta_3)$ indicates that i punishes j more, the farther j's contribution is below (above) the average level of others. In the estimation, each individual pair of players in the same group in a period is the unit of observation. As shown in Table 4, the coefficient β_2 is positive and significant in all three treatments, for both the data for all periods and for period one alone. This indicates that agents receive more punishment, the less they have contributed relative to others in the group. The negative coefficients on β_3, which are significant in two of the three treatments, indicate that the more an individual contributes relative to others, the lower the punishment he receives. The effect of others' average contributions, as captured with β_1, is ambiguous in sign. While

[9] Highly similar results are obtained if the average contribution including player j is used in the specification instead of the average excluding player j. This is also the case for the data presented in Table 5, as well as when the average contribution including player i is used in the estimation reported in tables 6a and 6b rather then the average omitting player i. The results are available from the authors.

Table 4 Sanctions assigned by i to j in second stage as a function of contribution decisions of j in stage one of current period

$$p_i^{j2t} = \beta_0 + \beta_1 \bar{c}_{-i}^{\,t} + \beta_2 \max\left\{0, \bar{c}_{-j}^{\,t} - c_j^t\right\} + \beta_3 \max\left\{0, c_j^t - \bar{c}_{-j}^{\,t}\right\} + \beta_4 t$$

	All periods			First period		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant	−3.174***	−2.512***	−0.229	0.321	−0.912	4.174***
	(0.353)	(0.251)	(0.324)	(0.932)	(0.882)	(1.130)
Others' average contribution (β_1)	0.014	0.054***	−0.192***	−0.187**	−0.070	−0.606***
	(0.021)	(0.017)	(0.021)	(0.091)	(0.071)	(0.127)
Amount recipient contributed below average (β_2)	0.387***	0.165**	0.438***	0.215**	0.238*	0.389***
	(0.034)	(0.023)	(0.028)	(0.088)	(0.136)	(0.084)
Amount recipient contributed above average (β_3)	−0.103*	−0.038	−0.135***	−0.032	−0.201	−0.245***
	(0.053)	(0.034)	(0.042)	(0.092)	(0.145)	(0.083)
Period (β_4)	−0.168***	−0.116***	−0.107***			
	(0.021)	(0.015)	(0.017)			
Log-likelihood	−1123.95	−1729.53	−1263.27	−147.92	−174.16	−167.94
Observations	2880	2880	2880	144	144	144

***1% significance level, **5% significance level, *10% significance level, Tobit estimation used. Standard errors are in parentheses

the effect is significantly positive in the Revenge Only treatment, it is significantly negative in the No Revenge treatment.

In the third stage, there are several potential motivations for sanctioning. Agents may wait until the third stage to sanction low contributors,[10] they may enforce sanctions that others failed to apply in stage two, or they may counterpunish. We consider the influences of each of these effects in an estimation of Eq. (6). Table 5 contains the estimates from the following regression model for the Full Information and the No Revenge treatments:

$$p_i^{j3t} = \beta_0 + \beta_1 p_j^{i2t} + \beta_2 \left\{ \left(\sum_{m \neq j} \sum_{k \neq i,j} p_m^{k2t} \right)/2 \right\} + \beta_3 \bar{c}_{-i}^{\,t}$$
$$+ \beta_4 \max\left\{0, \sum_{k \neq i} p_j^{k,2t} - \left(\sum_{m \neq j} \sum_{k \neq i,j} p_m^{k2t} \right)/2 \right\}$$

[10] The convexity of the cost function for punishment in each stage means that there are cost savings from spreading out punishment allocations over the two stages. This property, in principle, might encourage a greater quantity of total punishment. Previous studies indicate that agents punish more, the lower the price of punishment (Anderson and Putterman 2005; Carpenter 2005a, b; Casari 2005). However, the fact that there is less punishment in the three treatments with two punishment stages than in the Baseline treatment indicates that other forces more than offset any such effect.

$$+\beta_5 \max\left\{0, (\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t})/2 - \sum_{k\neq i} p_j^{k,2t}\right\}$$
$$+\beta_6 \max\left\{0, \bar{c}_{-j}^{t} - c_j^t\right\} + \beta_7 \max\left\{0, c_j^t - \bar{c}_{-j}^{t}\right\} \quad (6)$$

The dependent variable in Eq. (6) is the number of punishment points that player i assigns to player j in the third stage of period t. The coefficient β_1 takes on a positive value if counterpunishment occurs. The coefficient is positive if player i reciprocates sanctions he receives by assigning more punishment points to j in stage three, the more points j assigned to i in stage two of the same period. The variable $\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t}/2$ is the average number of punishment points assigned to individuals other than i and j in stage two. The variable $\sum_{k\neq i} p_j^{k2t} - (\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t})/2$ is the difference between the total number of punishment points that j assigned to individuals other than i and the average number of punishment points assigned to individuals other than i and j in stage two. The coefficient β_4 is positive if i sanctions j more, the more punishment that j has disbursed to players other than i, relative to the average punishment. If β_5 is positive, sanction enforcement is occurring, since it means that the fewer points j assigns relative to the average punishment of third parties in stage two, the more i sanctions j in stage three. The coefficients β_6 and β_7 capture the dependence of sanctioning behavior in stage three on contribution decisions in stage one, and if punishment of relatively low contributors occurs in stage three, β_6 is positive. The variables indicating the average punishment of third parties and j's deviation from the average are not included in the Revenge Only treatment. This is because subjects cannot calculate the number of punishment points j has assigned and the relevant average from the information they have available.

The estimates, reported in Table 5, show that counterpunishment, sanction enforcement, and stage three punishment of low contributors all occur. The coefficients β_1 on the variable indicating the number of sanctions assigned in the second stage are positive and significant in all three treatments, indicating the existence of counterpunishment, applied with increasing severity as the initial sanction is increased. This pattern is also consistent with a pattern of blind vengeance in the No Revenge treatment. Even in the No Revenge treatment, when individuals are not aware of who has sanctioned them, they apparently use information about the sanctions others receive. They appear to conjecture that those who punish others more are also relatively likely to have punished them, and try to avenge the sanctions that they have received in stage two of the game by targeting these high punishers.

The table also shows that sanction enforcement occurs. Players receive more punishment in stage three, the fewer sanctions they assign in stage two compared to the average punishment level assigned. The evidence is the significantly positive coefficients on β_5 in the Full Information and No Revenge treatments, although the effect is only borderline significant for the period 1 data. Thus, failure to punish low contributors in stage two with the severity others view as appropriate draws punishment in stage three. The significantly positive coefficients on β_6 in the Full Information and the No Revenge treatments, indicate that low contributions in stage one are also punished in stage three, as they are in stage two.

Table 5 Number of punishment points that player i assigns to j in the third stage as a function of prior contribution and sanctioning decisions of recipient

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

	All periods			First period		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant	−4.603*** (0.352)	−3.778*** (0.336)	−2.129*** (0.376)	−0.423 (0.814)	−2.307** (0.934)	−0.409 (0.699)
Points j assigned to i in second stage (β_1)	0.520*** (0.106)	1.151*** (0.096)	0.659*** (0.088)	0.528* (0.299)	1.243*** (0.227)	0.517*** (0.122)
Others' average punishment in second stage(β_2)	0.014 (0.111)		−0.416*** (0.148)	−0.230 (0.367)		−0.335* (0.175)
Others' average contribution (β_3)	0.194*** (0.022)	0.015 (0.018)	−0.084*** (0.019)	−0.120 (0.085)	−0.007 (0.068)	−0.148** (0.065)
Positive deviation of recipient from average punishment in second stage (β_4)	0.219* (0.119)		−0.618*** (0.179)	0.610 (0.391)		−0.386* (0.205)
Negative deviation of recipient from average punishment in second stage (β_5) (Sanction Enforcement)	0.405*** (0.104)		0.298*** (0.101)	0.679* (0.403)		0.255* (0.145)
Amount recipient contributed below the average (β_6) (Punishment of Low Contributors)	0.170*** (0.028)	0.002 (0.029)	0.286*** (0.027)	0.179*** (0.060)	−0.531** (0.239)	0.177*** (0.061)
Amount recipient contributed above the average (β_7)	−0.060 (0.042)	−0.053 (0.038)	−0.012 (0.037)	−0.547*** (0.196)	−0.110 (0.102)	−0.006 (0.049)
Period (β_8)	−0.129*** (0.020)	−0.041** (0.017)	−0.068*** (0.017)			
Log-likelihood	−1403.59	−1043.78	–979.69	–87.84	–64.34	−106.19
Observations	2880	2880	2880	144	144	144

Table 6 The effect of period t sanctions on changes in contribution between periods t and $t+1$: low contributors

$$c_i^{t+1} - c_i^t = \beta_0 + \beta_1 \sum_k p_k^{i2t} + \beta_2 \sum_k p_k^{i3t} + \beta_3 \left(c_i^t - \bar{c}_{-i}^t\right)$$

	All periods			Period 1		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant ($\beta 0$)	0.556***	−1.887***	0.4633**	2.874***	1.606	0.643
	(0.71)	(0.1693)	(0.1863)	(0.856)	(1.572)	(0.480)
Points received in second stage of period t (β_1)	0.6109***	0.4353***	0.2198***	1.216***	1.562***	0.852***
	(0.0734)	(0.0921)	(0.0943)	(0.346)	(0.160)	(0.234)
Points received in third stage of period t (β_2)	−0.1461*	0.2239**	0892	−0.859***	0.123	0.903***
	(0.0832)	(0.1168)	(0.1324)	(0.312)	(0.619)	(0.212)
Deviation from others' average contribution in period t (β_3)	−0.2785***	−0.8374***	−0.37***	0.200	0.472*	−0.339***
	(0.0365)	(0.0326)	(0.05)	(0.189)	(0.278)	(0.100)
R^2	0.170	0.47	0.182	0.158	0.626	0.562
Observations	1,182	1,218	828	84	72	78

***1% significance level, **5% significance level, *10% significance level, Standard errors are in parentheses

3.3 The effect of sanctions

The relationship between points received in each of the two stages of period t and subsequent contributions in period $t+1$ is described in Tables 6 and 7, in which the results of the following estimation are presented.

$$c_i^{t+1} - c_i^t = \beta_0 + \beta_1 \sum_k p_k^{i2t} + \beta_2 \sum_k p_k^{i3t} + \beta_3 \left(c_i^t - \bar{c}_{-i}^t\right). \tag{7}$$

In this equation, the dependent variable is the change in the contribution of individual i between period t and $t+1$. A positive value of the dependent variable indicates that contributions increase from one period to the next. The coefficient β_1 captures the effect of the number of punishment points received in stage two of period t on the change in contributions. If β_1 is positive, it indicates that the receipt of a larger quantity of sanctions has the effect of inducing a greater subsequent net increase in contributions. While stage two sanctions are presumably unambiguously interpreted as punishment for contribution decisions, stage three sanctions, as we have seen, may reflect other motivations. The extent to which stage three sanctions change subsequent contributions is captured with the coefficient β_2. The deviation from the others' average contribution, whose effect is captured with β_3, is included as an explanatory variable to account for any regression to the mean in contributions that is independent of the number of sanctions received. Such regression to the mean may reflect a desire to conform to the average contribution. It

Table 7 The effect of period t sanctions on changes in contribution between periods t and $t+1$: high contributors

$$c_i^{t+1} - c_i^t = \beta_0 + \beta_1 \sum_k p_k^{i2t} + \beta_2 \sum_k p_k^{i3t} + \beta_3 \left(c_i^t - \bar{c}_{-i}^t\right)$$

	All periods			Period 1		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant ($\beta 0$)	0.5023***	0.4438***	1.0967***	−1.704	2.120*	2.971**
	(0.1689)	(0.1650)	(0.1836)	(1.982)	(1.118)	(1.177)
Points received in second stage of period t (β_1)	0.1052	−0.0725	−0.0939	−1.525	−0.697**	−0.498
	(0.1132)	(0.096)	(0.1020)	(1.001)	(0.313)	(0.309)
Points received in third stage of period t (β_2)	0.1966**	−0.7655***	0.2115	1.584	−0.577	−0.552
	(0.1015)	(0.133)	(0.1609)	(1.003)	(0.390)	(0.995)
Deviation from others' average contribution in period t (β_3)	−0.6758***	−0.6219***	−0.459***	−0.540*	−0.489***	−0.484***
	(0.0356)	(0.035)	(0.036)	(0.323)	(0.121)	(0.115)
R^2	0.23	0.209	0.171	0.063	0.206	0.307
Observations	1,218	1,380	840	60	72	54

***1% significance level, **5% significance level, *10% significance level, Standard errors are in parentheses

may also be the result of pure randomness: an individual who independently draws a contribution level each period would exhibit regression toward his mean contribution. Table 6 shows the estimates for *low contributors*, those who contribute less than the group average in period t, while Table 7 gives the same data for *high contributors*, who contribute more than the group average in period t. The data are separated into high and low contributors because previous work suggests that these two groups may react differently to the receipt of sanctions (Masclet et al. 2003).

The estimates show that low contributors who receive more punishment points in stage two of period t, respond with a more positive net change in contributions for period $t+1$. The coefficient β_1 is significantly positive at the 1% level in all three treatments. Punishment has the intended effect of inducing low contributors to increase their contributions in the next period. However, the same is not the case for high contributors, for which none of the β_1 coefficients is significantly positive. The β_2 coefficients show no general pattern for either high or low contributors, suggesting that receiving sanctions in stage three is not interpreted as punishment for low contributions. For both high and low contributors, the β_3 coefficient is significantly negative in all three treatments, revealing the existence of a general tendency of regression to the mean in contribution levels. The higher one's contribution relative to the average, whether it is above or below, the stronger the tendency is to lower it in the following period.

Sanction enforcement leads individuals to increase the quantity of sanctions that the recipient assigns in stage two of the following period, while counterpun-

Table 8 The effect of stage three punishment on sanctions assigned in the second stage of following period: low punishers

$$\sum_k p_i^{k,2,t+1} - \sum_k p_i^{k,2,t} = \beta_0 + \beta_1 \sum_k p_k^{i3t} + \beta_2 \left(\sum_k p_k^{i2t} - \overline{\sum_k p_k^{2t}} \right)$$

	All periods			Period 1		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant	0.3508*** (0.1300)	0.2245*** (0.0545)	0.1906 (0.0614)	0.130* (0.072)	−0.750 (0.618)	0.275 (0.211)
Points received in third stage of period t	0.1536** (0.0655)	0.1278* (0.069)	0.0796** (0.0308)	0.188*** (0.062)	0.115 (0.159)	1.313*** (0.158)
Deviation from Average punishment in period t	−0.1109 (0.2029)	−0.0400 (0.0840)	0.0717 (0.0979)	0.008 (0.067)	−3.115** (1.288)	0.858* (0.464)
R^2	0.012	0.005	0.010	0.141	0.183	0.466
Observations	498	720	660	60	30	90

ishment reduces the quantity assigned. The effects appear in Tables 8 and 9, which display the results of the estimation of Eq. (8), for low and high punishers, respectively. A low (high) punisher in period t is an individual who distributed fewer (more) punishment points in stage two of period t than the average in her group. The mean sanction assigned in stage two is only discernable to individuals in the Full Information and the No Revenge treatments.

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

The dependent variable in the equation is the change in the total amount of punishment that player i assigns between stage 2 of period t and stage 2 of period $t + 1$. The independent variables are the total number of points the individual has received in stage three of period t and the difference between the number of points he assigns and the average number of points individual members of the group assign in stage two of period t. If $\beta_1 > 0$, subjects respond to sanction enforcement or to counterpunishment with increases in the quantities of sanctions they assign in stage 2 of the following period. If $\beta_2 < 0$, there is a tendency for those who have sanctioned less relative to the average in stage two of period t, to exhibit a greater net increase in the sanctions they assign in stage two of period $t + 1$ relative to stage two of period t.

The estimates show that in the Full Information and No-Revenge treatments, the greater the number of sanctions a low punisher receives in stage three of period t, the greater the net increase in the number of punishment points he distributes in stage two of period $t + 1$ relative to period t. He acts as if he has interpreted the punishment he has received as sanction enforcement, and responds as if to reduce the receipt of future sanction enforcement. No such effect is observed for high punishers, who do not appear to interpret stage three sanctions they receive as punishment for insufficient assignment of sanctions. Under the Revenge Only

Table 9 The effect of stage three punishment on sanctions assigned in the second stage of following period: high punishers

$$\sum_k p_i^{k,2,t+1} - \sum_k p_i^{k,2,t} = \beta_0 + \beta_1 \sum_k p_k^{i3t} + \beta_2 \left(\sum_k p_k^{i2t} - \overline{\sum_k p_k^{2t}} \right)$$

	All periods			Period 1		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant	0.5445***	0.6735***	0.3246	1.601***	2.184***	0.604
	(0.11505)	(0.1128)	(0.145)	(0.329)	(0.514)	(0.415)
Points received in third stage of period t	0.875	−0.4840***	−0.0325	−0.085	−0.531	−1.342***
	(0.0556)	(0.0553)	(0.0829)	(0.199)	(0.460)	(0.357)
Deviation from Average punishment in period t	−0.9997***	−0.4510***	−0.6184***	−1.152***	−0.815***	−0.723***
	(0.0367)	(0.0364)	(0.0418)	(0.204)	(0.233)	(0.072)
R^2	0.5839	0.3415	0.272	0.422	0.747	0.666
Observations	564	690	600	66	48	54

***1% significance level, **5% significance level, *10% significance level, standard errors are in parentheses

treatment, low punishers cannot be identified, and stage three sanctions are interpreted as counterpunishment. Consequently, in Revenge Only, the more sanctions that one receives in stage three of period t, the fewer one assigns during stage two of period $t + 1$. This effect is observed for both low and high punishers for the data for all 20 periods, though not for the data from $t = 1$. Thus, the use of counterpunishment in stage three for prior sanctions has the effect of deterring the sanctioner in the next period.

After the above effects are taken into account, the more that individuals punish in excess of the average sanction in the second stage of a given period, the greater the tendency to sanction less in the following period. This effect is observed in all three treatments for high sanctioners, as can be seen from the negative and significant β_2 coefficients in each of the three treatments for high punishers. No general tendency toward (or away from) conformity is detected for low punishers.

3.4 The six stage full information treatment

The decrease in contributions, which is borderline significant, that the introduction of a second stage of punishment induces suggests that additional stages of unrestricted punishment might further reduce contributions. Figure 1 indicates that average contributions are lower throughout the time horizon of the sessions in 6SFI than in the Baseline treatment. The results of a Mann–Whitney rank sum test for treatment differences, using each group's average contribution for a session as the unit of observation, given in Table 2, reveal that the differences in median group contributions between 6SFI and Baseline are significant at the 5% level.

The effect on welfare of the additional stages of punishment opportunities is large and negative. Figure 2 illustrates the effect, displaying both mean and median

welfare for 6SFI. In the early periods of the 6SFI treatment, average earnings are negative. In later periods it increases, but remains below the level in all other treatments, except for the Revenge Only treatment. Indeed, average individual earnings remain below 20 over the entire session, indicating that earnings are lower than at a benchmark where contributions are zero from all players and no punishment is possible. Figure 2 also shows the earnings of median individual in each period, which display a similar pattern as the mean, indicating that the low earnings are not due to very low earnings on the part of a few individuals. As Table 2 indicates, the welfare level over the entire 20 periods is significantly lower in 6SFI than in each of the other four treatments.

The source of the lower welfare is twofold. While contributions are lower in the 6SFI than in the Baseline and the Full Information treatments, the number of sanctions applied is also higher in 6SFI. Table 3 displays the data, indicating that the average number of sanctions an individual receives in a period is equal to 4.11 points. This represents a reduction of 41.1% of first stage earnings, not including the costs the sanctioners incur. The number of points assigned is 2.71 times the amount in the next highest treatment. The number of points assigned within a period follows a distinct pattern of small declines during stages 2–5, and a large increase in stage 6. Some of this activity in the last punishment stage appears to consist of sanctions for earlier contribution or punishment decisions that have been deferred until they are immune from counterpunishment.[11]

4 Conclusion

In this study, we investigate the impact of allowing punishment of sanctioning behavior on contributions, sanctioning decisions, and payoffs of groups facing a social dilemma. The results of our study show that the existence of multiple rounds of sanctions, in which any player may sanction any other, has a negative effect on the level of contributions relative to a setting with one round of sanctions. Our Six Stage Full Information treatment yields significantly lower contribution and welfare levels than our Baseline treatment with only one round of sanctions. Indeed, in this treatment, average welfare is lower than at a benchmark in which no sanctioning mechanism at all is present and contributions are zero for each individual.

Our treatments with two rounds of sanctions allow us to measure the relative magnitudes of the effects on contributions of opportunities to engage in counterpunishment and sanction enforcement. Our results are consistent with those reported in the recent work of Nikiforakis (2004), who finds that when agents are permitted to counterpunish, but not to enforce sanctions, contributions to the public good

[11] In some periods, a phenomenon of escalating counterpunishment is observed. This phenomenon consists of a sanction that player i applies to j, followed by the assignment of counterpunishment by j to i, and one of more reciprocal reprisals. We give two examples of this phenomenon here. In period 3, players A, B, C, and D in group 1 contribute 12, 8, 12, and 0 tokens, respectively. Player A then assigns 5 points to D in stage two. D responds by assigning 1 point to A in stage three. Then A assigns 3 to D, D assigns 2 to A, and A assigns 2 to D in the next three stages. Similarly, in period 6, the contributions of players A–D are 12, 10, 5, and 15 tokens respectively. In stage two, D allocates one point to C and C responds by assigning D one point in stage three. D assigns C two points, C gives one point to D, and D directs two points to C, respectively, in rounds 4–6.

decrease. Although contributions increase when sanction enforcement is introduced, the effect is not significant. This result is in line with that obtained in Cinyabuguma et al. (2004, 2005). We find that the increase in contributions from sanction enforcement is smaller in magnitude and only partially offsets the reduction in contributions due to counterpunishment. The overall effect of a second stage of punishment and full observability of prior contribution and punishment decisions is a (borderline significant) reduction in contributions, as the effect of counterpunishment on contributions is larger than the effect of sanction enforcement. As suggested by the data from the 6SFI treatment, additional rounds of sanctioning opportunities appear to further erode contribution levels.

The sanctions operate in an intuitive manner at the individual level. Agents sanction low contributors in the second stage. In the third stage they sanction low contributors and low sanctioners, as well as counterpunish. Sanctions received in the second stage increase recipients' contributions in the following period. Counterpunishment reduces the quantity of sanctions recipients assign in the following period, while sanction enforcement increases it.

As is well known, in the absence of a sanctioning mechanism, voluntary contributions are highly susceptible to the free-rider problem. On the other hand, the environment of our Baseline treatment or of Fehr and Gaechter (2000), with a single stage of sanctioning, is highly conducive to cooperation. The setting with five unrestricted opportunities to punish that we have studied here generates contribution levels that lie between those in these two extreme cases, but welfare levels that are lower than when no sanctioning system is available. Thus, sanctioning systems appear to be most effective in promoting cooperation when punishers are anonymous, in the sense that they are immune from the consequences of their sanctioning behavior, as in the setting of Fehr and Gaechter. The high levels of cooperation and observed in their environment appear not to be robust to the removal of this immunity. When sanctioning behavior can be punished, counterpunishment is common. This reduces welfare both because it is costly in itself and because it reduces contributions through its deterrence of punishment of low contributors. These effects are magnified when there are many rounds of punishment opportunities, as in our 6SFI treatment. In this treatment, costly episodes of reciprocal counterpunishment have a strong negative effect on welfare, and average contributions are relatively low. Permitting individuals to punish those who fail to sanction low contributors, or those who do sanction high contributors, reveals a reluctance to do so. Our results suggest that in a setting with repeated opportunities to target individuals for punishment, limiting opportunities for counterpunishment is welfare improving.

We believe that the major patterns in the data can be reconciled with game-theoretic analysis under appropriate assumptions on preferences. Under the classical assumption that individuals' payoffs are a function of only their monetary earnings, there is a unique subgame perfect equilibrium in the game. No punishment occurs in the last stage of the game, for any previous history of play. Because punishment in the next-to-last stage is costly to the sanctioner, and does not effect decisions at later nodes, no punishment is applied in the next-to-last stage for any previous history of play. Analogous backward induction arguments show that the unique subgame perfect equilibrium is that no contributions are made and no punishment is assigned at any time.

However, suppose instead that the game is one of incomplete information. Assume that there exist types of player who receive utility not only from their own income, but also for reciprocation. One type of player receives utility from reducing the earnings of those who take actions to lower his payoff. Another type receives utility from reducing the earnings of those who fail to take actions that uphold the social norm. These types of individuals would be willing to lower their monetary earnings to some extent to punish those who submit low contributions or those who fail to punish free riders, as well as to engage in counterpunishment. Our data are consistent with the existence of these types, since such punishment behavior is widespread in our data.[12]

If the above types of players exist, the game is one of incomplete information in which some individuals are income maximizers, some have the preferences described above, and individuals' types are private information. Consider the Full Information treatment. We conjecture that sequential equilibria exist with the following basic structure. In the last stage, some individuals, who receive utility from reciprocation, punish low contributors, engage in counterpunishment, and enforce sanctions, while other individuals, who are income maximizers, do not punish. In the preceding first punishment stage, some individuals punish free riders, either because they receive utility from doing so, and/or because they know there is a probability that they will be punished in the next stage should they not do so. Whether or not, and how much, an individual punishes depends on her type and her beliefs about the proportions of types in the population. In the contribution stage, some individuals may contribute in order to avoid punishment in the next stage, others would not contribute, and whether and how much an individual contributes depends upon his own type and his beliefs about the proportion of types. Indeed, there may exist multiple equilibria and the multiplicity may account for the heterogeneity among groups. We believe that a complete game-theoretic analysis of this game with the above assumptions on preferences, while beyond the scope of the current project, would be a valuable tool in explaining the data from this and other recent experimental studies on the effect of punishment on cooperation.

References

Anderson, C., Putterman, L.: Do non-strategic sanctions obey the law of demand? the demand for punishment in the voluntary contributions mechanism. Games Econ Behav (forthcoming) (2005)

Andreoni, J.: Why free ride: strategies and learning in public goods experiments. J Public Econ **35**(1), 57–73 (1988)

Bochet, O., Page, T., Putterman, L.: Communication and punishment in voluntary contribution experiments. J Econ Behav Organ (forthcoming) (2005b)

Bolton, G., Ockenfels, A.: ERC: A theory of equity, reciprocity, and cooperation. Am Econ Rev **90**, 163–193 (2000)

Carpenter, J.: Punishing free riders: how group size affects mutual monitoring and the provision of public goods. Games Econ. Behav. (forthcoming) (2005a)

Carpenter, J.: The demand for punishment. J Econ Behav Organ (forthcoming) (2005b)

Casari, M.: On the design of peer punishment experiments. Exp Econ **8**(2), 107–115 (2005)

[12] See for example Bolton and Ockenfels (2000) or Dufwenberg and Kirchsteiger (2004), for development of models of reciprocal preferences.

Cinyabuguma, M., Page, T., Putterman, L.: On perverse and second-order punishment in public goods experiments with decentralized sanctioning. working paper. Providence: Brown University (2004)
Cinyabuguma, M., Page, T., Putterman, L.: Can second order punishment deter perverse punishment. working paper. Providence: Brown University (2005)
Coleman, J.: Foundations of Social Theory. Belknap Press of Harvard University Press (1990)
Dufwenberg, M., Kirchsteiger, G.: A theory of sequential reciprocity. Games Econ Behav **47**, 268–298 (2004)
Falk, A., Fehr, E., Fischbacher, U.: Driving forces of informal sanctions Econometrica (forthcoming) (2005)
Fehr, E., Gaechter, S.: Cooperation and punishment in public goods experiments. Am Econ Rev **90**(4), 980–994 (2004)
Fischbacher, U.: z-Tree: A toolbox for readymade economic experiments. working paper. Zurich: University of Zurich, Institute for Empirical Research in Economics (1999)
Gaechter, S., Fehr, E.: Collective action as a social exchange. J Econ Behav Organ **39**(2), 341–369 (1999)
Gaechter, S., Herrmann, B.: Norms of Cooperation Among Urban and Rural Dwellers: Experimental Evidence from Russia, mimeo. Harvard University and the University of Cambridge, Nottingham (2005)
Isaac, R.M., McCue, K., Plott, C.: Public goods provision in an experimental environment. J Public Econ **26**(1), 51–74 (1985)
Isaac, R.M., Walker, J.: Group size effects in public goods provision: the voluntary contributions mechanism. Q J Econ **103**(1), 179–199 (1988a)
Isaac, R.M., Walker, J.: Communication and free-riding behavior: the voluntary contributions mechanism. Econ Inquiry **26**(4), 585–608 (1988b)
Ledyard, J.: Public goods: a survey of experimental research. In Kagel, J. and Roth, A. (eds.) Handbook of Experimental Economics, Princeton: Princeton University Press, pp. 111–194
Masclet, D., Noussair, C., Tucker, S., Villeval, M.: Monetary and non-monetary punishment in the voluntary contributions mechanism. Am Econ Rev **93**(1), 366–380 (2003)
Nikiforakis, N.S.: Punishment and Counter-punishment in Public Goods Games: Can We Still Govern Ourselves? working paper. Royal Holloway: University of London (2004)
Noussair, C., Tucker, S.: Combining monetary and social sanctions to promote cooperation. Econ Inquiry **43**(3), 649–660 (2005)
Ostrom, E., Walker, J., Gardner, R.: Covenants with and without a sword: self-governance is possible. Am Polit Sci Rev **86**(2), 404–17 (1992)
Yamagishi, T.: The provision of a sanctioning system as a public good. J Personality Soc Psychol. **51**(1), 110–116 (1986)