The Journal of Socio-Economics 40 (2011) 919–928

Contents lists available at SciVerse ScienceDirect

The Journal of Socio-Economics

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

Sanctioning as a social norm: Expectations of non-strategic sanctioning in a public goods game experiment

Jana Vyrastekova[a,*], Yukihiko Funaki[b], Ai Takeuchi[c]

[a] Nijmegen Center for Economics (NiCE), Nijmegen School of Management, Radboud University Nijmegen, The Netherlands
[b] Waseda University, Department of Economics, Tokyo, Japan
[c] Graduate School of Economics, Waseda University, Tokyo, Japan

ARTICLE INFO

Article history:
Received 1 November 2010
Received in revised form 20 June 2011
Accepted 29 August 2011

JEL Classification:
C7
C9
H4

Keywords:
Non-strategic sanctions
Unobserved sanctions
Social norm of sanctioning
Public goods
Economic experiment

ABSTRACT

Sanctioning increases cooperation in public goods games, but not indiscriminately under all conditions and in all societies, and the mechanisms by which sanctioning exercises its impact on behavior are yet to be studied in detail. We show experimentally that in the presence of sanctioning, our experimental subjects adjust their behavior in order to avoid being a free rider. They do this not only in the STANDARD sanctions treatment, where they directly experience any sanctions assigned to them, but also in our main treatment, the SECRET sanctions treatment, where no information on sanctions received is available until the end of the experiment. We observe no such free riding avoidance in the treatment without sanctioning. The mere knowledge that sanctions might be assigned increases cooperation among the members of our subject pool; subjects expect that non-strategic sanctioning occurs against the free riders. Moreover, these expectations are correct as we observe a similar pattern and extent of sanctioning in both treatments. We propose that sanctioning in itself is a social norm and may be culturally dependent, as suggested in the literature.

1. Introduction

Free riding in social dilemmas can be prevented if players are able to assign costly sanctions to their co-players. This is the case in cultures that sanction free riders.[1] In this study, we assert that sanctioning is a social norm. It is a rule that prescribes which situations merit sanctions, and it is accompanied by a set of beliefs that correctly predict punishable situations. In our experiments, we find that subjects avoid the free rider position in a public goods game and that their beliefs about the punishable behavior are correct since the actual sanctioning is targeted towards the free riders. This happens even in cases where the sanctions are not observed, i.e., when no evidence about the actual sanctions assigned is available to the subjects during the experiment. Moreover, there seems to be little attrition in the sanctioning expectations over time. In our experiments with unobserved sanctions, subjects contributing less than others on an average increase contributions in the next period. This prevents the unravelling of cooperation; and instead, gives rise to group-specific norms. These group-specific cooperation levels are driven by initial contributions, and by contribution strategies that seek to avoid being seen as a free rider.

Several pieces of evidence in the literature have emphasized the role of the subjects' home-grown beliefs about sanctioning. For example, in experiments in which free riders are sanctioned, subjects respond to the introduction of sanctioning options into the experiment in a way that is consistent with the actual sanctioning behavior. Consequently, contributions to the public goods increase immediately after the announcement of the opportunity for costly sanctions (e.g. Fehr and Gächter, 2002). Fehr and Fischbacher (2004) find that stated beliefs reveal that subjects expect a third-party punishment to be imposed on unfair dictators. Moreover, one-shot dictators become more generous under the "threat" of receiving verbal commentary on their distributional decisions (Ellingsen and Johannesson, 2008). An interesting finding on sanctioning is the existence of a multiplicity of sanctioning norms across societies. Gächter and Herrmann (2007, 2009) study societies in which cooperators are sanctioned along with free riders. They find that this type of sanctioning is paired with a decrease in cooperation rates when the sanctioning opportunities are added

* Corresponding author at: Nijmegen University, Department of Economics, Postbus 9108, 6500 HK Nijmegen, The Netherlands. Tel.: +31 24 36 11588; fax: +31 24 36 12379.
E-mail address: J.Vyrastekova@fm.ru.nl (J. Vyrastekova).

[1] These are usually Western societies, where the majority of experimental studies has been performed (see for example Ostrom et al., 1992; Fehr and Gächter, 2000; Masclet et al., 2003; Egas and Riedl, 2008; Anderson and Putterman, 2006; van Soest and Vyrastekova, 2006; and Carpenter, 2007).

1053-5357/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.socec.2011.08.020

to the experiment, as opposed to the increase in contributions that is normally observed in studies with sanctioning of free riders. In such a society, subjects do not increase their cooperation, as they correctly anticipate that such behavior will be punished. This piece of evidence motivates our assertion that sanctioning behavior in a society is accompanied by a corresponding set of local beliefs consistent with the behavior.

Recent evolutionary approaches suggest that beliefs about sanctioning play a central role in explaining the survival of cooperation through sanctioning. For some time, sanctioning of free riders stands as a strong candidate for resolving the puzzle of human cooperation (Fehr and Gächter, 2000). Negative emotions held toward free riders are hypothesized as the proximate mechanism supporting outwardly costly sanctions (Fehr and Gächter, 2002). At the same time, the ultimate evolutionary mechanisms of sanctioning are less obvious. In a society, punishers receive lower payoffs than non-punishing cooperators as soon as free riders invade the population, which creates pressure against the survival of the punishing cooperators. Group-selection models which focus on the interactions among members of small isolated groups, would allow for the survival of cooperation by sanctioning of free riders, but these models require unrealistic assumptions (such as small group interactions, and no migration) which are unlikely to have prevailed during the course of human evolution.

Models that respond to this criticism employ a range of approaches, most of which emphasize the importance of social norms — systems of rules and shared beliefs that have been adopted by a significant part of the population. Gardner and West (2004) show that even at the level of individual gene evolution, altruistic (or non-strategic) sanctioning can survive if individuals adjust their cooperation levels to the threat of punishment that is present.[2] A society in which sanctioning and a cooperative response to the threat of sanctioning co-evolve ends up with high levels of cooperation that do not require kin- or group-selection arguments.

Such a relationship between the threat of sanctioning and cooperation might be created if beliefs in sanctioning get transmitted in the population parallel to the sanctioning traits themselves. In a similar vein, Henrich and Boyd (2001) sustain cooperation by using altruistic sanctioning in a model with cultural group selection in which norms evolve within groups and are transmitted from more successful groups to groups with lower fitness. Finally, Gintis (2003) presents a gene-cultural evolution model, in which the individual genetic evolution process of traits for norm internalization is coupled with norm transmission between groups. In this way, the norm of sanctioning of free riders can arise as one of the possible stable states of the evolutionary process.

An important common element in these evolutionary approaches is the relationship between sanctioning, as an internalized norm or as an individual trait, and a system of beliefs that coevolves with it and that leads to the expectations consistent with the actual sanctioning behavior. One implication of these models is that individuals in societies where sanctioning supports cooperation through the sanctioning of free riders are expected to believe that sanctioning will be used to sanction free riders, and to believe this without empirical evidence on the use of sanctions. A second, perhaps more important implication, is that sanctions can affect behavior even when not being used. As a direct consequence, when sanctioning of free riders is internalized by a society as a social norm, the overall costs incurred in sanctioning might be much lower than usually assumed.

The experimental findings presented in this paper provide evidence that the impact of sanctions assigned to the free riders is driven by expectations rather than by any direct impact of sanctions. Subjects correctly predict that free riding will be sanctioned and hence adjust their behavior accordingly. Sanctioning thus is a form of a social norm—a behavioral rule that is adhered to and that is expected to be adhered to by a significant fraction of the population.

We obtain data on the role of the subjects' expectations regarding sanctioning behavior by varying the timing of the feedback given to the subjects about the sanctions they receive in a repeated public goods game. We implement two information treatments. In both of them, subjects first play several rounds of the linear public goods game without any sanctioning, and experience the convergence of their contributions to full free riding. After re-grouping subjects for the second part of the experiment, we allow them to assign costly sanctions to other group members in each round of the game. In the STANDARD treatment, subjects receive feedback on the sanctions assigned to them in the same round in which they made their contribution to the public good, while in the SECRET treatment, sanctions can be assigned in each round, but they are revealed to their receivers only at the end of the experiment. All strategic (forward-looking) motivations for sanctioning are thus removed in the SECRET treatment. Moreover, if behavior in the public goods game is affected by the presence of the sanctioning option in the SECRET treatment, then we ascribe this effect to the beliefs subjects hold about sanctioning behavior of others. Note that we do not explicitly elicit subjects' beliefs about sanctioning but infer them from the subject's behavior, and from comparison of our experimental treatments.

Our paper contributes to the existing literature on the origins of sanctioning, and the way sanctions affect behavior. Most closely, it links to the study by Fudenberg and Pathak (2010). Using a random matching design, they compare one-shot public goods games with immediately observed sanctioning to games with unobserved sanctions, in order to differentiate between repeated-game theory motivations for sanctioning, and altruistic (backwards-looking) sanctioning. The authors report that subjects are more likely to punish, and they punish harsher when sanctions are not observed, but no explanation is offered for this effect. The wide-spread use of unobserved sanctioning has been documented in other studies as well. Abbink et al. (2004), for example, compare immediate and delayed feedback on ultimatum offer rejections in a random matching design, and report a considerable rejection rate in the covered response design, going beyond the effect of creating a group reputation for toughness in order to eliminate unfair offers. Non-strategic sanctioning in a one-shot public goods game is observed by Walker and Halloran (2004). In their true one-shot experiment, sanctions do not disappear, although it seems that subjects are not able to anticipate correctly the level of contribution they will be sanctioned for. Note that so far, no conclusion can be made about the impact the unobserved sanctions would have on behavior over time, and whether unobserved sanctioning is a temporary or lasting phenomenon. In our experiments, we therefore study a repeated public goods game, which allows us not only to analyze the origins of sanctioning, but also the way subjects incorporate the threat of the (unobserved) sanctioning into their contributions in a public goods game.

In brief, we find that expectations of sanctioning enforce cooperation. Our data corroborates the widespread use of non-strategic (backward-looking) sanctions in public goods games. Sanctioning occurs in both of our experimental treatments, and similar levels of free riding in the public goods game are punished to a similar

[2] In the literature, sanctions are referred to as altruistic if they are costly to the sender, but yield no benefits to him/her (Fehr and Gächter, 2002). The sanctioning studied in this paper is most precisely referred to as non-strategic, meaning that the sanction providers engage in sanctioning without planning to benefit from it, despite the fact that they might end up better off with than without the sanctioning option available. However, both altruistic and non-strategic sanctions are used non-strategically, without their choice being based on the expected benefit from the sanctioning decision.

extent, no matter whether the sanctions are announced immediately to the sanctioned individuals or whether no such information is provided. Strikingly, cooperation in groups with unobserved sanctions does not unravel, but these groups develop group-level specific contribution over time, when responding to the contributions of others, increasing contributions if others contribute more, and decreasing them otherwise. Sanctioning thus proves to have a lasting effect on the levels of cooperation in a public goods game even without having an immediate effect on the welfare of the sanctioned individuals. We conclude that sanctioning system operates as a social norm, and comes accompanied by beliefs of what situations merit punishment.

The remainder of this paper is organized as follows. In Section 2, we describe the game and the experimental design. The data are analyzed in Section 3, and Section 4 concludes.

2. The game and experimental design

2.1. The game

We implement a four-player repeated public goods game in which the payoff function for the ith player, $i = 1, .., 4$, is given by:

$$\pi_i = 10 - x_i + 0.4(x_i + X_{-i}) \quad (1)$$

Here, the contribution x_i that an individual i makes to the public good comes from a set of integers $x_i \in \{0, 1, \ldots, 10\}$, and X_{-i} is the sum of the contributions of all players other than i.

This public goods game is extended by including a sanctioning stage. After each round, players observe the contribution vector $\{x_1, \ldots, x_4\}$, receive an additional endowment of $S = 10$, and can assign any integer number from that endowment to any of the other players. Let us denote by s_{ij} the amount of points player i assigns to player $j, j \neq i$, where $s_{ij} \in \{0, 1, \ldots, 10\}$ and $\sum_{j \neq i} s_{ij} \leq 10$. Player i's payoff from sanctioning in one round is given by:

$$f_i = 10 - \sum_{j \neq i} s_{ij} - 3\sum_{j \neq i} s_{ji} \quad (2)$$

Player i's total payoff in one round of the game with sanctioning is thus given by

$$\Pi_i = \pi_i + f_i. \quad (3)$$

The personal benefits from contributing to the public good are lower than the benefits of not contributing; therefore, according to the backwards induction argument, all players will choose zero contributions to the public good in all rounds. However, this solution is inefficient because in this linear public goods game, the total payoff is maximized if each group member chooses the maximal contribution of 10. The same prediction can be made for the case when sanctioning is present, since sanctions are costly to the sender while yielding no direct material benefit. Consequently, rational and payoff-maximizing players will not use any sanctions, and the contributions to the public good when sanctioning is possible will be equal to the contributions in a public goods game without sanctioning.

This prediction, however, is likely to be behaviorally irrelevant in societies where non-strategic sanctioning of free riders occurs. Non-strategic sanctioning is driven by different motivations than simply by evaluating the expected material outcomes. Instead, the roots of non-strategic sanctioning may be found in the emotional processes of humans, e.g., in the pleasure of taking revenge or the desire to uphold norms such as those of fairness. Hopfensitz and Reuben (2009) found that self-reported anger was related to sanctioning decisions (see also Fehr and Gächter, 2000), and Falk et al. (2005) provided evidence of frequent norm-driven sanctioning. Quervain et al. (2004) offered direct neurological evidence on the processes underlying sanctioning; that is, they measured brain activation patterns and related sanctioning to an anticipated gratification. The fact that sanctioning is related to pleasurable emotions may explain why (some) subjects are willing to incur the material costs of sanctioning. Generally speaking, non-strategic sanctioning has been observed frequently in such experiments, and it has been found to play a bigger role than the strategic sanctions assigned under the expectation of a direct material benefit (Falk et al., 2005; Casari and Luini, 2009; Walker and Halloran, 2004; Fudenberg and Pathak, 2010).

In our experiments, we use two different types of sanctioning treatments. In the STANDARD sanctions treatment, subjects contribute to the public goods game, observe the individual contributions of others in their group, and are allowed to decrease the payoffs of other players by assigning deduction points. Each deduction point decreases the payoff of the person receiving the deduction point by three points. All subjects learn how many deduction points they received and are informed of the final payoffs.

In our main treatment, the SECRET sanctions treatment, we capture the effect of beliefs in non-strategic sanctioning. It differs from the STANDARD sanctions treatment only in that it delays the feedback on the sanctions received (if any) until the end of the experiment. This arrangement removes the strategic (repeated-game theoretic, forward-looking) motivations of sanctioning. In this treatment, contributions to the public good can be affected only by the subjects' beliefs in non-strategic sanctioning and in its specific form that is expected to be operational in the population.

Our predictions can be summarized by the following two hypotheses. The first hypothesis relates to the type of non-strategic sanctioning expected to be used by our sample population, i.e., the sanctioning of free riders. The second, main hypothesis concerns the subjects' expectations on what constitutes behavior that exposes a person to the possibility of receiving a sanction, as revealed through the changes in the individual contributions to the public good. In a population where free riders are sanctioned, this dynamic is expected to be driven by the avoidance of being seen as a free rider. The fact that this type of sanctioning is a norm implies that the same type of behavior will be observed with and without direct evidence on sanctioning. We thus state the following hypotheses:

Sanctioning hypothesis: Sanctioning is non-strategic: thus, the probability and the size of the sanctions assigned is the same for both the SECRET and the STANDARD sanctions treatment. The form of the sanctioning is the sanctioning of free riders.

Expectations of sanctioning hypothesis: Subjects believe that sanctioning will be targeted toward free riders in the group. Consequently, subjects adjust their contributions in repeated interactions over time so as to avoid being a free rider in the group. This is the case both for the SECRET sanctions treatment and the STANDARD sanctions treatment.

2.2. Experiment design

In the fall of 2006, we conducted four experimental sessions at Tilburg University in the Netherlands. A total of 64 subjects participated in either the STANDARD or the SECRET treatment, with each subject randomly assigned to only one treatment (see summary in Table 1). The participants were bachelor's and master's degree students of economics, law, and business. The language of the experiment was English. Upon arrival, participants were randomly assigned to a computer terminal, and were informed that the experiment consisted of two tasks; the instructions for each task

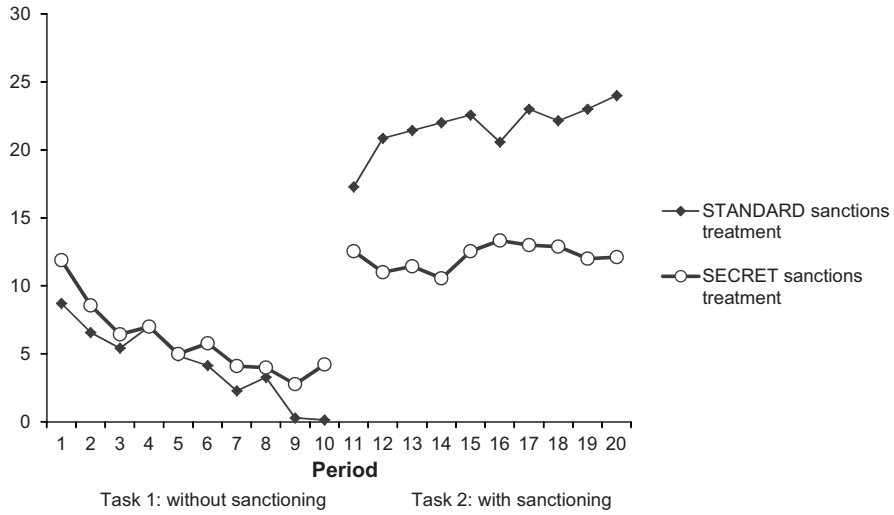

Fig. 1. Average group contributions to the public good per treatment and period.

Table 1
Summary of all sessions.

Session	Treatment	Periods 1–10	Periods 11–20	Number of subjects
1	SECRET	Without sanctions	With unobservable sanctions	20
2	SECRET	Without sanctions	With unobservable sanctions	16
3	STANDARD	Without sanctions	With observable sanctions	12
4	STANDARD	Without sanctions	With observable sanctions	16

were read aloud just before the relevant task started. The experiment was fully computerized and the software was programmed using z-Tree (Fischbacher, 2007).

In Task 1 of the experiment, after being arranged into groups of four, the subjects participated in 10 rounds of the repeated public goods game without any sanctioning. The subjects received both the payoff function and the payoff table. They were informed that they would stay in the same group for all 10 rounds and that in each round the subjects would be assigned a new integer label ranging from 1 to 4. This matching protocol allowed the subjects to share knowledge about the behavior of others in the group across rounds, while severing the relationship between decisions made in the current round and the behavior of a specific individual in the past. In each round, each subject received 10 tokens and was asked to divide them between option I (public good) and option II (private good). All formulations were neutral. After each round, each subject observed the contributions of all other subjects in his or her group.

After round 10, at the beginning of Task 2, we regrouped the subjects into new groups and informed them that they would stay in the new group for all 10 rounds of Task 2. The subjects' labels were changed after each round, as in Task 1. Each round of Task 2 had two stages. In Stage 1, the subjects chose their contributions to the public good, they observed the contributions of other individuals, and they received 10 tokens for the second-stage endowment. In Stage 2, each subject was given the chance to assign any of his or her Stage 2 endowment to any of the other subjects in the group. Each assigned point decreased the payoff of the receiver by three points.

In the STANDARD sanctions treatment, we informed the subjects how many sanctioning points they received during the same round in which these points were assigned to them. In the SECRET sanctions treatment, by contrast, all information on sanctions received was delayed until the end of the experiment. After all 10 rounds of Task 2, each subject learned how many Stage 2 sanctioning points he or she received in each round.

The experiment lasted about 1.5 h, and participants earned an average of 10,70 euros (including a 3-euro participation fee).

3. Data analysis

We start by summarizing contributions to the public good in both treatments. We then explain contribution behavior by studying evidence on sanctioning under observable and unobservable sanctions, and by retrieving beliefs about sanctioning from the dynamics of the contributions.

3.1. Contribution behavior

Previous research has convincingly shown that observable sanctions, where one sanction unit is three times as costly to the receiver as to the sender, serves as an effective tool for supporting cooperation. This is also the case in our experiments with the STANDARD sanctions treatment (see Fig. 1). Average group contributions are clearly higher when subjects are able to sanction (Task 2, rounds 11–20) than when sanctions are not possible (Task 1, rounds 1–10) (Mann–Whitney test with a unit of observation being the average group contribution in Task 1 and in Task 2, $N = 14$, $p = 0.026$).[3] Subjects achieve more than 25% of the maximum group contributions (equal to 40) in the SECRET sanctions treatment, and more than 50% of the maximum in the STANDARD sanctions treatment. Compared to previous studies, our STANDARD sanctions treatment seems rather standard in the sense that (i) contributions are well above half of the endowment, on average, and (ii) they increase rather than decrease over time.[4]

[3] Recall that we re-grouped subjects between Task 1 and Task 2.

[4] Using the same 1:3 sanctioning technology as we do, Gächter et al. (2008) report contributions close to ours in their 10-round repeated game. These authors show that the longer the time horizon, the more likely the contributions increase towards full efficiency with this technology. Note that in their seminal paper, Fehr and

At first sight, the impact of unobserved sanctions on behavior is less obvious. Average group contributions in the SECRET sanctions treatment do not differ between Task 1 and Task 2 (Mann–Whitney test with a unit of observation being the average group contribution in Task 1 and in Task 2, $N = 18$, $p = 0.730$). This might suggest that subjects in the SECRET sanctions treatment do not take into account the fact that free riders will be sanctioned.

However, there is also no significant difference in the *average group contributions* in Task 2 between the two treatments, STANDARD and SECRET sanctions treatment (Mann–Whitney U test with a unit of observation being the average group contribution in Task 2 of the STANDARD treatment and in Task 2 of the SECRET sanctions treatment, $N = 16$, $p = 0.210$), nor is there a difference between the payoffs that individuals earned in the Task 2 in the two treatments (Mann–Whitney U test with a unit of observation being the average individual payoff in Task 2 of the STANDARD and in Task 2 of the SECRET sanctions treatment, $N = 16$, $p = 0.133$). We need to look beyond treatment averages to understand these observations and the effect that unobserved sanctions might have.

To this end, the evolution of group contributions over time in the two sanctioning treatments is revealing (see Fig. 2). This evolution explains why, on average, there is no difference in group level cooperation between the STANDARD and SECRET sanctions treatments. We observe a strong bifurcation of cooperation levels in the STANDARD sanctions treatment (see panel (c) in Fig. 2), with groups converging either to full cooperation (4/7) or to full free riding (3/7). On the other hand, in the SECRET sanctions treatment, the evolution of contributions shows a different dynamic (see panel (d) in Fig. 2), with half of the groups converging to full defection (4/9) and the other half (5/9) preserving or increasing the initial contribution levels over time. The absence of an end-game effect or of a negative time trend in these groups is remarkable and offers evidence that unobservable sanctions can have a long-term impact on cooperation. Focusing on individual groups reveals that the contribution dynamics differs fundamentally when sanctions are observable and when sanctions are not immediately observable. Moreover, the presence of sanctions, though unobservable, results in higher contributions, as can be seen from comparing panel (b) and panel (d) of Fig. 2.

In summary,

Observation 1: Contributions to the public good are higher in the presence of sanctions than in the absence of sanctions, both in SECRET and STANDARD sanctions treatment.

3.2. Sanctioning behavior

Above, we established that unobserved sanctions affect contributions in at least half of the groups in our experiments, and their impact extends over all periods of interaction. Before studying the mechanisms by which the unobserved sanctions affect behavior, we now first ask in what way – and if at all – do subjects use them, in particular in comparison to the use of sanctions in the STANDARD sanctions treatment.

In the STANDARD sanctions treatment, subjects receive information on the total number of sanctioning points assigned to them by their co-players in every round. In this case, the early rounds investments into sanctioning in a repeated public goods game could be compensated by future benefits if the sanctioned

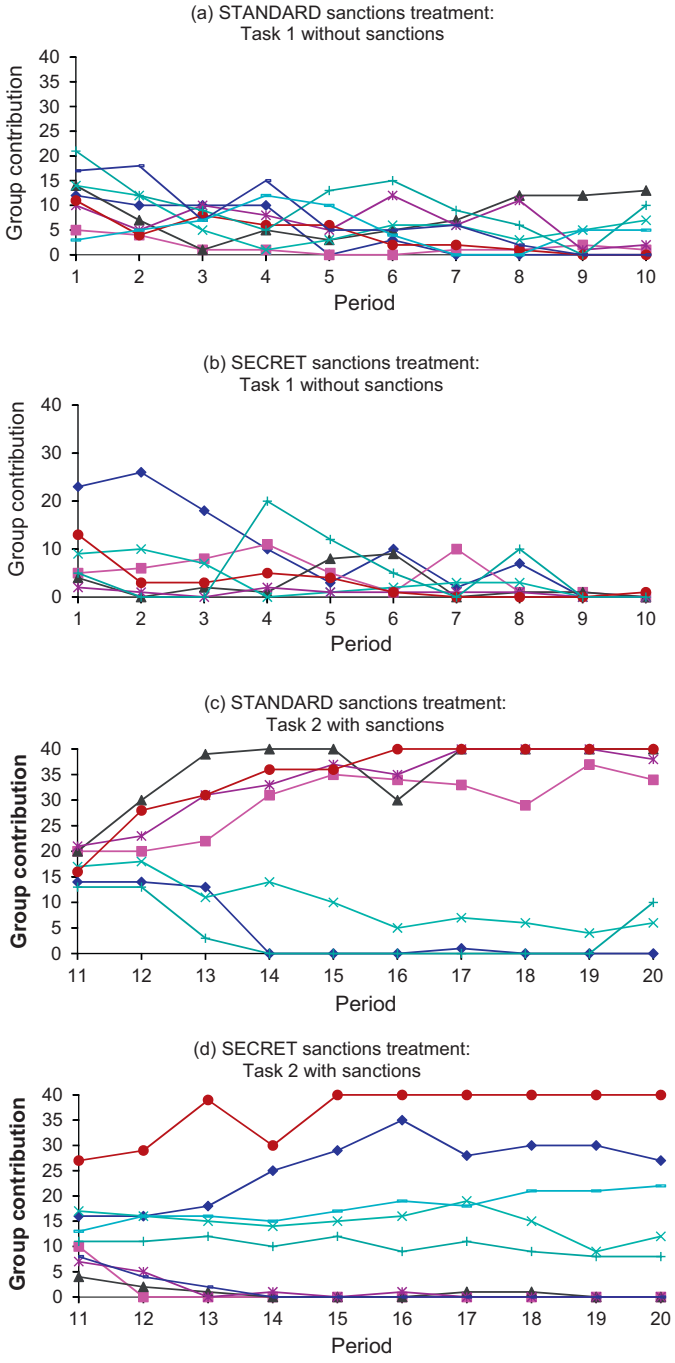

Fig. 2. Group contributions per treatment.

individuals increase their contributions to the public good after receiving sanctions. Subjects, disregarding their preference for sanctioning itself, therefore might have incentives to sanction free riders in the STANDARD sanctions treatment in order to establish reputation as free rider intolerant individuals. This is not the case in the SECRET sanctions treatment, which allows assigning of sanctions in every round but does not provide subjects with any feedback on the sanctions they received until the end of the experiment. Thus, in the SECRET sanctions treatment, strategic (forward-looking) incentives for sanctioning aimed as deterring free rider's behavior are absent. The only sanctioning incentives in the SECRET sanctions treatment are thus non-strategic (backwards-looking).

Gächter (2000) use a different sanctioning technology, where each unit of sanction assigned decreases the payoff of the receiver by 10%. This implies that sanctioning of free riders is mostly more effective than when using a 1:3 sanctioning as in our paper; and is accordingly more effective at achieving full cooperation in a public goods game.

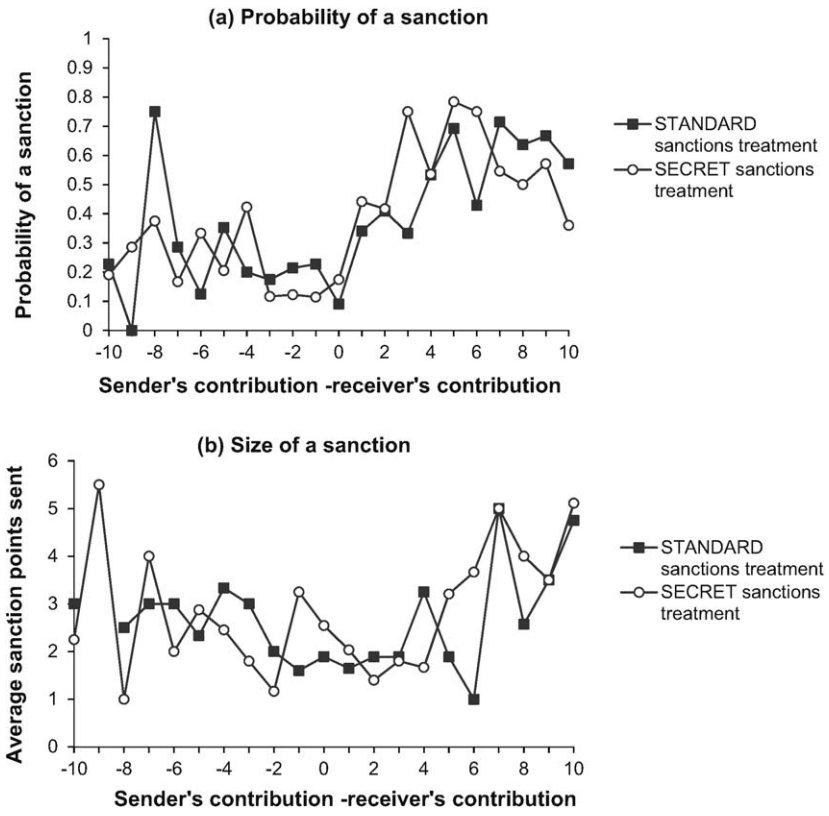

Fig. 3. (a) Probability of sanction and (b) size of a sanction (assuming sanction is assigned) as a function of the difference between the sender's and the receiver's contributions.

Nevertheless, we find that both the probability of assigning a sanction and the extent of the sanctioning in a given situation do not differ in our two treatments, which is a strong evidence that the motivations behind the sanctioning in our experiment are mostly non-strategic. Fig. 3 summarizes the evidence, with panel (a) illustrating the probability that an individual assigns any sanctioning points and panel (b) showing the average number of sanctioning points conditional on assigning a sanction. The independent variable in these figures is the difference between the sender's and the receiver's contributions to the public good. To interpret the figure, note that in the second half of the experiment, about 90% of all individual observations fall into the categories where the difference between the contributions of the sanctioning and the sanctioned individual is between −5 and 5.

In our study, we find that cooperators sanction free riders: the sanctions are mostly targeted toward receivers who contribute less than the sender. Some spiteful sanctioning, targeted at receivers who contribute more than the senders, is found as well (see also Falk et al., 2005). Conditional upon choosing to sanction another subject, the sanctioning subject assigns on average equivalent sanctions for equivalent degrees of free riding in the two treatments, as defined by the difference between the sender's and receiver's contribution to the public good. Furthermore, consistent with non-strategic explanations of sanctioning, the subjects assign sanctions over all periods of the SECRET sanctions treatment, including the last period of the STANDARD sanctions treatment (see Fig. 4).

The seemingly prevailing differences in the average sanctioning across treatments over time, see Fig. 4, can be understood by noticing that the contributions differ over time in the two treatments. In order to address properly whether the sanctioning likelihood and pattern do not differ across treatments, we need to apply a regression analysis controlling for the contribution behavior in a group. We do that by estimating a hurdle model, allowing for the sanctioning decision to be guided by a process different from the decision on how much to sanction. Sender i's decision to sanction receiver j is explained by a logit model (see column 2 in Table 2). The size of the sanction assigned by i to j is estimated by a truncated Tobit model (see column 3 in Table 2). Individual random effects are taken into account in both parts of the model. The treatment dummy variable secret was set to 1 in the SECRET sanctions treatment and to 0 in the STANDARD sanctions treatment.

The regression results confirm that free riders are punished by cooperators, and that the decision to sanction is equally likely in both treatments (see significant coefficient of the variable $\max(x_i - x_j, 0)$ in column 2 in Table 2, and an insignificant interaction term with the treatment variable). When a subject assigns a sanction, its size does not depend on the treatment, as supported by the economically very small and insignificant interaction terms (see column 3 in Table 2), but it does depend on the extent of free riding. Sanctioning is not only statistically but also economically relevant in the sense that every point of a difference between the

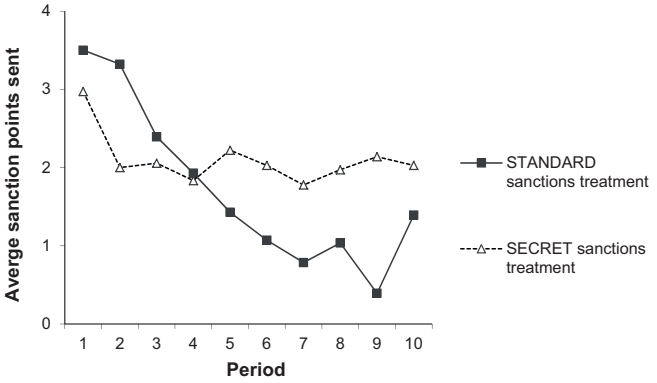

Fig. 4. Average sanction points sent per period.

Table 2
Regression analysis of subject *i*'s decision to sanction or not sanction subject *j* (in column 2), and of the size of sanction assigned by subject *i* to subject *j*, assuming a decision to sanction was made (in column 3). Column 2 contains the average marginal effects of the independent variables of a logit model with individual random effects. Column 3 contains the coefficients of a truncated Tobit model with individual random effects. Standard errors in parentheses.

	Decision to sanction	Size of sanction
Secret	−0.016	−0.781
	(0.086)	(0.880)
$\max(x_j - x_i, 0)$	−0.008	−0.205**
	(0.006)	(0.97)
$\max(x_j - x_i, 0) * \text{secret}$	0.006	0.038
	(0.009)	(0.139)
$\max(x_i - x_j, 0)$	0.042**	0.363***
	(0.008)	(0.076)
$\max(x_i - x_j, 0) * \text{secret}$	0.0005	0.046
	(0.008)	(0.113)
Period	−0.027***	−0.021
	(0.007)	(0.094)
Period * secret	0.012*	0.194*
	(0.007)	(0.114)
Constant		0.219
		(0.737)
N	1920	488
Log likelihood	−905.62028	
Wald χ^2	174.64	56.98

* **p** < 0.1.
** **p** < 0.05.
*** **p** < 0.001.

contributions of the sanctioning sender and the sanctioned receiver is accompanied by an implied payoff decrease of more than one point (equal to 0.363 * 3). Effectively, senders assign sanctions in a way that eliminates the payoff advantage of a free rider in the contribution stage of the game. These payoff losses experienced by free riders are independent of the treatment. The regression analysis thus reveals that sanctioning is triggered by motivations present in both treatments, i.e., motivations not linked to the subject receiving a feedback on being sanctioned. Finally, note that there is some negative time trend in sanctioning. The period coefficient is negative when estimating the probability of assigning a sanction (amounting to about 10% decrease in sanctioning probability per four periods). This decrease is present in both treatments, though. We do not find any significant time trend when estimating the size of the sanctions.

In view of these results, we offer the following observation:

Observation 2: Sanctions are mainly non-strategic. When the differences in the sender's and receiver's contributions to the public good are taken into account, subjects sanction free riders with equal probability and equal strength in both the STANDARD and SECRET treatments.

3.3. Contribution dynamics

How do subjects react to the observed and unobserved sanctions in the repeated public goods game? Without evidence on sanctioning in the SECRET sanctions treatment, subjects have to rely on their experience with sanctioning norms, and hence on their home-grown beliefs they bring into the laboratory. To address these beliefs, we first comment on the individual's contribution strategies and then present a regression analysis of the contribution dynamics. It is our goal to demonstrate that subjects can assess correctly what kind of behavior elicits sanctions, by observing how subjects change contributions based on the feedback received.

The evidence on the sanctioning behavior in our experiments shows that free riders, by contributing less than others, are exposed to receiving sanctions, both observable and unobservable. Thus, an individual who correctly understands sanctioning behavior should be aware of this and avoid the free rider position in the group. Do subjects really behave this way? To address the type of contribution adjustment strategies individuals use, we identify for each individual, how he or she changes own contribution in the next round depending on how his or her contribution compares to the average contribution of others in the group in the current round. In particular, we obtain information for each individual on how he or she updates own contributions in cases (i) when contributing less than others on average in the group, and in cases (ii) when contributing more than others on average in the group in the current round. In general, subject can either increase, not change, or decrease own contribution in the upcoming round. Note that we are only able to describe an individual's strategy if he or she ends up to be the below-average contributor in some rounds, and the above-average contributor in other rounds. Therefore, in this analysis we focus on subjects for who we do have information on behavior in both cases, (i) and (ii). Only a few subjects' strategies cannot be fully described (either because they always contributed less than others, or always contributed more than others on average), and we present here data on the remaining 61 subjects in Task 1 (rounds 1–10, without sanctioning) and 55 subjects in Task 2 (rounds 11–20, with sanctioning). Several strategy types are used frequently by our subjects: (1) a static strategy (no change of contribution over time), (2) a two-sided reciprocal strategy (increasing the contribution if below average and decreasing the contribution if above average), and (3) a positive-only reciprocal strategy (increasing the contribution if below average but no change if above average). Fig. 5 shows the population distribution of these strategies.

When sanctioning is not possible (in Task 1), the subjects most frequently apply the static strategy and the two-sided reciprocal strategy. When explicit sanctions are available (in Task 2), subjects do not anymore use covert sanctioning via decreasing their own contributions (as they often do when sanctions are not available, in Task 1) but they can chose an overt negative reciprocal action and do not decrease their contributions as cooperators.[5] This is the case in both sanctioning treatments, although more so in the STANDARD sanctions treatment. In the SECRET sanctions treatment, the static strategy is more popular than the positive reciprocal strategy, but also in this treatment, subjects use negative reciprocation via contributions less frequently when sanctioning is available than when sanctions are not available. Interestingly, the positive reciprocal strategy becomes more popular in both sanctioning treatments in comparison to the Task 1 without sanctioning. These changes in the strategies used in both treatments demonstrate the impact of the subjects' beliefs in sanctioning.

To explain the contributions adjustments in a detail, we run a Tobit regression analyzing the decision of subject *i* to change his or her contribution to the public good across two periods, $x_{i,t} - x_{i,t-1}$, see Table 3.[6] The explanatory variables are the distance of *i*'s contribution from the average contribution of others if contributing more than average in the previous period, or if contributing less than average in the previous period; these are the variables $\max(x_{i,t-1} - \bar{x}_{-i,t-1}, 0)$ and $\max(\bar{x}_{-i,t-1} - x_{i,t-1}, 0)$, respectively. We interact these variables with a dummy variable. In column 2, the dummy variable captures the comparison of Task 1 with Task 2 in the SECRET sanctions treatment (*dummy* = 1 in

[5] Until now, the decline in contributions in the public goods experiments has been mostly interpreted in terms of a rational best response by conditional cooperators towards the population containing free riders (Burlando and Guala, 2005). Our individual analysis view suggests that a part of the unravelling in public good provision may also be attributed to a covert sanctioning behavior.

[6] The changes in subject *i*'s contributions are truncated at 0 and at the endowment constraint, but these bounds are never binding. Including individual random effects has no impact on the results.

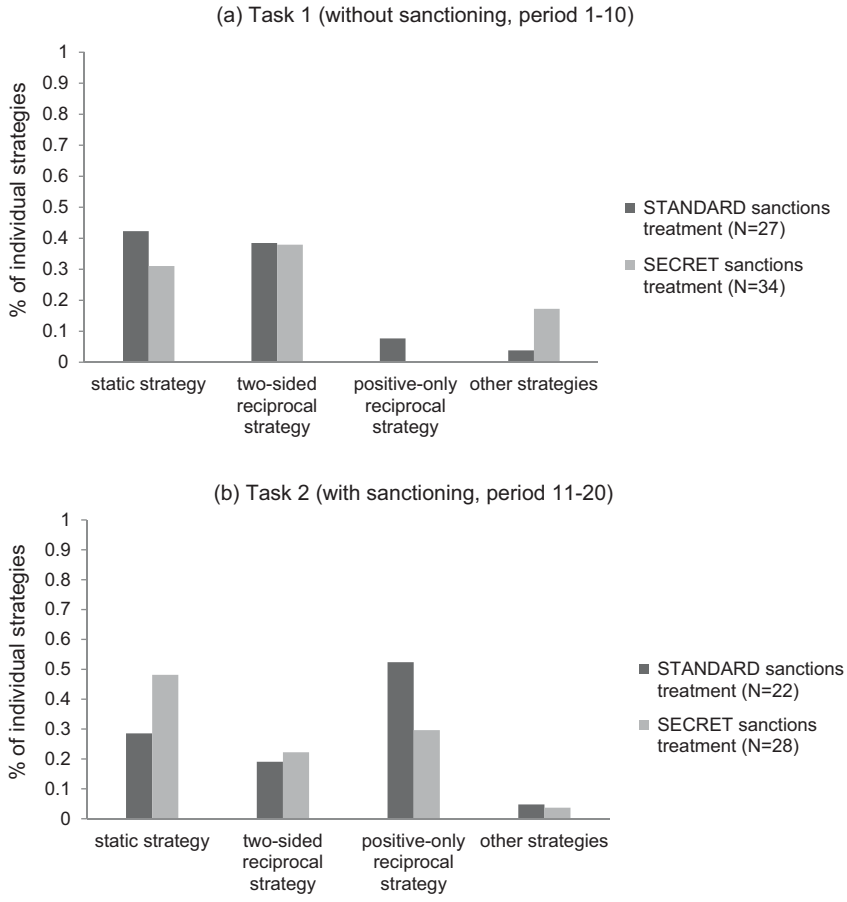

Fig. 5. Individual strategies (a) with and (b) without sanctioning available.

periods 1-10, and *dummy* = 0 in periods 11–20). In column 3, the dummy captures the comparison of Task 2 in the SECRET sanctions treatment with Task 2 in the STANDARD sanctions treatment (*dummy* = 1 in the periods 11–20 in the SECRET sanctions treatment, and *dummy* = 0 in the periods 11–20 in the STANDARD sanctions treatment).

We find that subjects respond differently to their relative position in the group, in terms of the contributions made to the public good, when the sanctioning option is not available (Task 1 of the SECRET sanctions treatment) and when the sanctioning option is available (Task 2 of the SECRET sanctions treatment). This is true despite the fact that in both cases the subjects receive feedback

Table 3
Explaining the change in an individual *i*'s contribution between period $t-1$ and period t by a censored Tobit model with individual random effects. Standard errors in parentheses.

Treatment/data	Secret/rounds 1–10 vs.11–20 (i) Dummy = rounds 1–10	Standard vs. Secret/rounds 11–20 (ii) Dummy = secret
Dummy	0.247	−1.889**
	(0.446)	0.805
$\max(x_{i,t-1} - \bar{x}_{-i,t-1}, 0)$	−0.203***	−1.267***
	(0.064)	0.139
$\max(x_{i,t-1} - \bar{x}_{-i,t-1}, 0) * \text{dummy}$	−0.339***	0.366*
	(0.082)	0.189
$\max(\bar{x}_{-i,t-1} - x_{i,t-1}, 0)$	0.360***	0.263**
	(0.008)	0.128
$\max(\bar{x}_{-i,t-1} - x_{i,t-1}, 0) * \text{dummy}$	−0.219**	−0.020
	(0.107)	0.180
Period	0.025	−1.048***
	(0.041)	0.089
Period * dummy	0.016	0.221*
	(0.059)	0.114
Constant	−0.290	8.699***
	(0.300)	0.638
N	640	640
Log likelihood	−1319.9805	−1681.5973

* **p** < 0.1.
** **p** < 0.05.
*** **p** < 0.001.

only on the public goods contributions, and not on the use of the sanctions (if any). In particular, as can be seen in Column 2 of Table 3, unobserved sanctions lead to a smaller (less than half in size) decrease of the above-average contributions and to a greater (more than three times as high) increase of the below-average contributions, as compared with the situation in which the same subjects adjusted their behavior in the absence of the sanctioning option.[7] In this way, the presence of sanctions, even if their use is not observable, supports cooperation in Task 2.

We now study how subjects adjust their contributions based on how these compare to the contributions of others in their group. In particular, we compare whether and how these contribution adjustments differ across the treatments with observable and unobservable sanctions (see Column 3, Table 3). There are no significant interaction effects suggesting that the subjects increase their contributions to avoid being a free rider, independent of whether or not the sanctions are observable. The treatment dummy is significant, though, resulting in lower unconditional contributions in the SECRET sanctions treatment.

We conclude that the presence of sanctions, even when not directly observable, generates incentives for avoiding the free rider position and stabilizes the contributions of cooperators. We thus offer the following observation:

Observation 3: Subjects respond to the presence of sanctioning even when sanctions are unobservable. They exhibit stronger incentives for avoiding the free rider position and a weaker tendency to avoid the sucker's position when sanctions are possible than in the absence of any sanctioning.

4. Conclusions

The sanctioning of free riders in social dilemmas might be the answer to the puzzle of why human cooperation occurs in large groups of unrelated individuals. Such sanctioning is often observed in experimental studies and has been shown to have non-strategic origins. Costly sanctions are imposed regardless of whether the sanctioning individual can count on benefiting from the action in the future. Recent evolutionary models suggest that costly sanctioning behavior in large populations, which would appear to decrease fitness, can be explained by taking into account the role of social norms in the process. When behavior is affected by expectations of free riders being sanctioned, even without direct evidence of such sanctioning, the social costs of sanctioning can be kept low. Groups sharing the norm of sanctioning free riders could obtain an evolutionary advantage over groups sharing a different sanctioning norm.

In our study, we find evidence that in a population in which sanctioning of free riders occurs, the subjects do indeed hold correct beliefs about the type of behavior that will result in being sanctioned. We make this observation based on comparing the behavior in a treatment in which sanctions are directly observable with a treatment in which the sanctions remain secret until the end of the experiment. In short, we find that subjects are equally likely to use the unobserved and observed sanctions in our experiment, and they punish similar instances of free riding with similar sanctions in the two treatments.

When costly sanctioning is a social norm, individuals' beliefs will shape behavior independent of the information on the actual sanctions received. This is precisely what happens in our experiments. When subjects know that sanctions are not available in their interactions, their behavior can be described as a reversion to the mean, with the subjects increasing contributions when below the average contribution of others, and decreasing contributions when above the average contribution. Introducing the possibility of assigning costly sanctions to this environment results in a lower tendency to decrease above-average contributions and a higher tendency to increase below-average contributions. Subjects' expectations on the sanctioning of free riders motivate them to avoid being a free rider, and in this way stabilizes cooperation at a group-specific level.

Our findings support the assumption built into the recent evolutionary explanations of costly sanctioning – that beliefs in costly sanctioning of free riders co-evolved with the individual traits for the sanctioning of free riders. The sanctioning of free riders can become a social norm shared by a population, although other forms of sanctioning norms are conceivable as well. This social-norm perspective on sanctioning may also explain why some societies are more cooperative than others. The level of cooperation in a society depends on the set of beliefs coevolving with the actual strategies supporting them. We might need to search further for the answer to the question what aspects shape these beliefs. Some indications can be found in the recent research (Herrmann et al., 2008), suggesting that culture, the shape of formal institutions and formal law enforcement interact with the informal enforcement rules found in a society.

Acknowledgements

We would like to thank the participants of the Tinbergen Institute seminar in Amsterdam and Charles Noussair for their valuable comments.

References

Abbink, K., Sadrieh, A., Zamir, A., 2004. Fairness, public good, and emotional aspects of punishment behavior. Theory and Decision 57, 25–57.
Anderson, Ch.M., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games and Economic Behavior 54, 1–24.
Burlando, R.M., Guala, F., 2005. Heterogeneous agents in public goods experiments. Experimental Economics 8, 35–54.
Carpenter, J.P., 2007. The demand for punishment. Journal of Economic Behavior and Organization 62, 522–542.
Casari, M., Luini, L., 2009. Cooperation under alternative punishment institutions: an experiment. Journal of Economic Behavior and Organization 71, 273–282.
Egas, M., Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. Proceedings of the Royal Society B 275, 871–878.
Ellingsen, T., Johannesson, M., 2008. Anticipated verbal feedback induces altruistic behavior. Evolution and Human Behavior 29, 100–105.
Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. Econometrica 73, 2017–2030.
Fehr, E., Fischbacher, U., 2004. Third party punishment and social norms. Evolution of Human Behavior 25, 63–87.
Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90, 980–994.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.
Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.
Fudenberg, D., Pathak, P., 2010. Unobserved punishment supports cooperation. Journal of Public Economics 94, 78–86.
Gardner, A., West, S.A., 2004. Cooperation and punishment, especially in humans. The American Naturalist 164, 753–764.
S. Gächter, B. Herrmann, The limits of self-governance in the presence of spite: experimental evidence from urban and rural Russia, CEDEX Discussion Paper (2007) 2007–11.
Gächter, S., Herrmann, B., 2009. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. Philosophical Transactions of the Royal Society 364, 791–806.
Gächter, S., Renner, E., Sefton, M., 2008. The long-run benefits of punishment. Science 322, 1510.

[7] To consider this, observe that one point of excess contribution above the group average motivates a 0.54 point decrease in contribution when sanctioning is not an option but only 0.20 point decrease when sanctioning is possible. Similarly, below-average contributors increase their contributions by 0.36 point when sanctions are an option, but only by one third of it, 0.14 point, without the sanctioning option.

Gintis, H., 2003. The hitchhiker's guide to altruism: gene-culture coevolution, and the internalization of norms. Journal of Theoretical Biology 220, 407–418.
Henrich, J., Boyd, R., 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of Theoretical Biology 208, 79–89.
Herrmann, B., Thöni, Ch., Gächter, S., 2008. Antisocial punishment across societies. Science 319, 1362–1367.
Hopfensitz, A., Reuben, E., 2009. The importance of emotions for the effectiveness of social punishment. Economic Journal 119, 1534–1559.
Masclet, D., Noussair, Ch., Tucker, S., Villeval, M.C., 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. American Economic Review 93, 366–380.
Ostrom, E., Gardner, R., Walker, J., 1992. Covenants with and without sword: self-governance is possible. American Political Science Review 86, 404–417.
Quervain, D.J.F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. Science 305, 1254–1258.
van Soest, D.P., Vyrastekova, J., 2006. Peer enforcement in common pool resource experiments: the relative effectiveness of sanctions and rewards, and the role of behavioral types. In: List, J. (Ed.), Using Experimental Methods in Environmental and Resource Economics. Edward Elgar.
Walker, J.M., Halloran, W.A., 2004. Rewards and sanctions and the provision of public goods in one-shot settings. Experimental Economics 7, 235–247.