Nicklisch, Andreas; Wolff, Ireneus

Working Paper

Cooperation norms in multiple-stage punishment

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2009,40

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Nicklisch, Andreas; Wolff, Ireneus (2009) : Cooperation norms in multiple-stage punishment, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2009,40, Max Planck Institute for Research on Collective Goods, Bonn

This Version is available at:
https://hdl.handle.net/10419/32237

**Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2009/40**

Cooperation norms in multiple-stage punishment

Andreas Nicklisch
Irenaeus Wolff

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2009/40

Cooperation norms in multiple-stage punishment

Andreas Nicklisch / Irenaeus Wolff

December 2009

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Cooperation norms in multiple-stage punishment*

Andreas Nicklisch
Max Planck Institute for Research on Collective Goods, Bonn
nicklisch@coll.mpg.de
and Irenaeus Wolff
University of Erfurt
irenaeus.wolff@uni-erfurt.de

December 8, 2009

Abstract

Carpenter and Matthews (2009) examine the cooperation norms determining people's punishment behavior in a social-dilemma game. Their findings are striking: absolute norms outperform the relative norms commonly regarded as the determinants of punishment. Using multiple punishment stages and self-contained episodes of interaction, we disentangle the effects of retaliation and norm-related punishment. An additional treatment provides data on the norms bystanders use in judging punishment actions. Our results partly confirm the findings of Carpenter and Matthews: only for the punishment-related decisions in the first iteration is the absolute norm outperformed by the self-referential norm set by the punisher's own contribution. For the decisions in all later iterations, as well as for bystanders' support in all iterations, the absolute norm organizes our data best. In contrast to the study by Carpenter and Matthews, we find an absolute norm of 3/4 of players' endowments to be both consistent across decisions and relatively stable over time.

Keywords: *Experiment, public-good, punishment, social norms, voluntary cooperation*
JEL-Classification: C92, D63, H41

*We are deeply indebted to Sophie Bade, Christoph Engel, Michael Kurschilgen, and Bettina Rockenbach for reading an earlier version of the paper and providing us with useful and detailed feedback. We would further like to thank the participants of the IMEBE workshop 2008 in Alicante for useful comments, and the Max Planck Society for financial support.

1 Introduction

Norms (i.e., common understandings about obligatory, permitted, or forbidden behavior)[1] influence our behavior in many real-world scenarios. People entering buildings keep doors open for others, parents' financial support for kindergarten initiatives is typically proportional to income – as we expect the tax burden to be – and men take their hats off when entering churches. There are numerous other examples of how norms guide behavior in groups, so that economics has devoted a substantial amount of effort to analyzing the influence of social norms in the last decades (important contributions include, e.g., Sugden, 1986, Sethi, 1996, or Sober & Wilson, 1998). Of particular interest for the economist's study of norms is their interplay with individual incentives. The archetype of a potential conflict between social norms and individual incentives is the social dilemma, where individual and collective interests are disaligned. Norm violations and others' responses to such violations have long been debated by the experimental literature in the context of decentralized sanctioning mechanisms. The latter have been shown to foster and maintain voluntary cooperation (seminal work has been provided by Ostrom et al., 1992, for common-pool resources, and Yamagishi, 1986, or Fehr & Gächter, 2000, for public goods). This paper sets out to analyze explicitly the norms of cooperation prevailing in situations of this kind, and systematically compares potential norm candidates in an experiment taylored to this purpose. More precisely, we elicit the norms employed in sanctioning uncooperative behavior when there are multiple sanctioning stages, and examine whether other group members who are not directly involved in the punishment actions share the same norms for sanctioning.

When thinking about cooperation norms in social-dilemma situations, one important distinction is that between relative and absolute norms. If a relative norm is made use of, the standard against which a player evaluates the behavior of others rises or drops along with the level of cooperation within the group. In other words, the cooperation level expected from an individual may be different in a cooperative group from the level expected in a less cooperative group. In contrast, absolute norms provide reference points for behavior independent of the group's current level of cooperation (for instance, there could be a norm always to cooperate fully). Relative norms have been estimated in a number of studies. Several authors rely on the average degree of cooperation within the group as the norm (Fehr & Gächter, 2000, 2002, Anderson & Putterman, 2006, and Sefton et al., 2007), while more recent studies focus on the degree of cooperation of the player

[1] Cf. Ostrom (2000).

who punishes (Herrmann et al., 2008, Egas & Riedl, 2008, Sutter et al., 2008, or Reuben & Riedl, 2009). Yet, little is known with respect to absolute norms and with respect to the question of whether relative or absolute norms guide cooperation and sanctions. An exception are Carpenter and Matthews (2009) who compare the predictive power of relative and absolute norms in explaining the sanctioning behavior. They show that by and large, absolute norms fit the data better than relative norms.

We extend the work of Carpenter and Matthews with respect to several important aspects. First, we are able to disentangle punishment related to a cooperative norm from acts of retaliation by (i) employing multiple sanctioning stages in conjunction with (ii) self-contained episodes of interaction (players change their interaction partners after each encounter). These features allow us to restrict counterpunishment actions to the individual episode of interaction, so that it does not directly affect the data obtained from later interactions. An interesting question following directly from the above is whether a persisting cooperation norm will play a role in higher iterations of punishment. Everyday experience tells us that the majority of situations share the feature of iterative punishment being possible. Experimental research has shown that behavior in such sequences can differ substantially from the behavior typically observed in simple settings of a single sanctioning stage (e.g., Denant-Boemont et al., 2007, Nikiforakis, 2008, and Nikiforakis & Engelmann, 2009).

The use of multiple sanctioning stages has a further advantage. It has long been known that a non-negligible fraction of punishment actions in social-dilemma situations is directed at high-contributors (e.g., Fehr & Gächter, 2000, or Cinyabuguma et al., 2006). However, to the best of our knowledge, none of the studies conducted on the topic has answered the question of what the motivations for such punishment are. The present paper makes a first step in that direction, being able to separate between retaliation and spiteful or competitive thinking. At the same time, we can largely rule out random errors as another possible source of high-contributor punishment suggested in the literature (ibd.).

On a second dimension, Carpenter and Matthews provide evidence that subjects employ different norms for the decisions of (i) whether to punish a player or not, and (ii) how hard they want to punish that particular player. We further explore this effect by *explicitly* disentangling both decisions: in our setting, players first announce to punish a certain player (at a cost), before deciding on the level of punishment in a second step. This will be interesting in a number of ways. It allows us to analyze the degree of consistency between the norms, both with respect to the question of whether the two decisions are triggered by relative or absolute norms, and that of whether

absolute norms – in case they matter – are similar across these decisions.

Finally, we provide additional insights on cooperation norms prevailing within groups by introducing an important treatment variation. In the standard setting, norms are revealed only indirectly by those players actively sanctioning others. However, there is a substantial number of players who abstain from punishment actions. Still, it is not clear whether this abstention is owed to the players' norms of cooperation not being violated, or whether it is due to other reasons, such as an aversion to forcing others by means of punishment, or that the costs of punishment are higher than the player's disutility from the norm violation. As far as these players' cooperation norm is concerned, the traditional setting provides little evidence. In order to elicit a cooperation norm using data from *all* players, we introduce a treatment condition in which, for each punishment action announced, those group members who are neither the punisher nor the punishee with respect to that specific action have to voice their (dis-)agreement with it. In order not to render the announced (dis-)approvals of players completely arbitrary, but to create some commitment with respect to these statements on norm-related behavior, all players are informed about them. As such, agreements and disagreements have no formal consequences, while they provide additional information on norms within a group. Further details concerning the experimental design are discussed in the follwing two sections.

Our results indicate that in line with the findings of Carpenter and Matthews, absolute norms seem to organize the decisions relating to norm violations very well. Particularly, we see that absolute norms unlike relative norms robustly predict punishment of norm violation across several punishment stages. Moreover, the data indicate that the absolute norm remains constant across various kinds of punishment decisions suggesting that approximately 3/4 of the maximum degree of cooperation serves as the prevailing cooperation norm, as long as it is not perturbed by retaliative actions. Interestingly, we can divide punishment stages into three categories. In the first stage, negative norm violation (i.e., cooperation levels of less than 3/4) triggers the announcement of and support for a punishment action and determines the severity of punishment. In the second stage, retaliation seems to enter as another main motivation for punishment. Finally, in the third stage, we observe a mixture of (counter-)retaliative actions and sanction enforcement, while the absolute cooperation norm remains at approximately 3/4 of the maximum level. The results suggest two things: there are persistent absolute norms for cooperation within small groups, while in stages 2 and 3, additional motives for punishment manifest themselves.

The remaining article is organized as follows: section 2 introduces the game and presents our research questions. Section 3 describes the experi-

mental design. Section 4 reports the results, while section 5 discusses the findings along with their implications.

2 The game and research questions

For our experimental investigation, we introduce two versions of a standard linear public-good game implementing a voluntary contribution mechanism with n players, $n \geq 2$, and multiple punishment stages, the BASIC game and the OPINION game. Both games consist of an endogenous (but finite) number of stages. In the *first step*, each player i receives an endowment of $e > 0$ monetary units and decides on her contribution x_i to the public good, with $0 \leq x_i \leq e$. Each monetary unit invested in the public-good has a marginal rate of per-capita return α, with $1/n < \alpha < 1$.

In the *second step*, each player is informed about the individual contributions to the public-good and the interim payoff which equals

$$\hat{\pi}_i = e - x_i + \alpha \sum_{j=1}^{n} x_j. \tag{1}$$

Furthermore, each player i announces whether and to which of the other players she wishes to assign punishment points. Punishment points $p_{i \to j}$ reduce the payoff of player j according to the details described below. Filing an announcement $a_{i \to j}$, $a_{i \to j} \in \{0, 1\}$, incurs a cost of $f_a > 0$ for i.[2]

In *step three*, the announcements are made public knowledge, and in our OPINION condition, the players being neither the punisher nor the target of an announcement $a_{i \to j}$, i.e., all players k s.t. $k \notin \{i, j\}$, may voice their opinion about the announcement. Opinions only take on one of two values, consent or dissent, and do not have any formal consequences for player i's action space and payoffs. Notice that without the previous announcement $a_{i \to j}$, player i is not allowed to assign punishment points to j under either treatment condition. In the BASIC condition, players are informed about all announcements, but cannot express their consent or dissent.

After players have voiced their opinions (if applicable), all players are informed about the number and the identity numbers of supporters in the *fourth step*. In this step, each player i simultaneously decides on the (integer) number of punishment points $p_{i \to j}$ she assigns at her private cost $c(p_{i \to j})$, where $p_{i \to j} \in [0, p^{max}]$. The punishment technology is such that each punishment point reduces the interim payoff of the punished player by ten percent,

[2] This procedure is designed to keep experimental subjects from announcing punishment actions "just in case" against every other subject.

and therefore, we have a natural limit for punishment points, $p^{max} = 10$.[3] Therefore, the payoff equals

$$\pi_i = \hat{\pi}_i \times \max\left\{0, (1 - 0.1 \sum_{j \neq i} p_{j \to i})\right\} - \sum_{j \neq i} c(p_{i \to j}) - F_a, \tag{2}$$

where F_a denotes the total number of announcements made by i times f_a and the cost function $c : \{0, 1, 2, ..., 10\} \mapsto \mathbb{R}$ is a strictly-monotone increasing function with $c(0) = 0$. All players are informed about the resulting payoffs.

If there has been at least one announcement to assign punishment points in step two, additional stages of steps 2 to 4 follow: we allow all players to make new announcements (each incurring costs of f_a). To avoid potential demand effects in the experiment, we do not impose a restriction of punishment opportunities to those who have been punished in the prior stage as, e.g., in the design of Nikiforakis (2008). Again, in the OPINION condition, players not directly affected by an announcement of player i against j simultaneously voice their opinion on the new announcements. New announcements allow players to increase the number of punishment points, even for players who have not been punished before.[4] All players are informed about the resulting payoffs. We repeatedly allow for new announcements and increases in punishment points until no player makes a further announcement to punish.[5] Notice that players can only apply for and execute further punishment if this does not cause their own current payoff π_i to become negative. Therefore, the number of iterations is finite and restricted at the most to $\sum_i \hat{\pi}_i / f_a$. Finally, players are informed about the payoffs and the game ends.

Since subjects play the game repeatedly over a finite number of rounds with changing anonymous interaction partners, the equilibrium of the game in both treatment conditions is rather obvious in light of standard theory according to which any player will only be concerned with his own monetary payoff. On the equilibrium path of the unique subgame-perfect equilibrium, nothing changes compared to the standard public-good game. If a player deviates making an announcement, other players are indifferent between endorsing and dissenting from the announced action. Whether it is endorsed or not, the player making the announcement does not have any incentive to carry out the punishment, as this is costly to her. Anticipating this, no

[3] We adopt the punishment mechanism already used by Fehr and Gächter (2000) and Nikiforakis (2008).

[4] Individual punishment costs are calculated according to the sum of points assigned per player, so that rationing the distribution of points across stages does not decrease costs.

[5] This procedure is similar to the one used by Nikiforakis and Engelmann (2009) in their multiple-stage treatments.

player will contribute to the public-good, since it is by $\partial\hat{\pi}_i/\partial x_i = -1+\alpha < 0$ a dominant strategy not to do so.

Thus, one can interpret contributions as voluntary cooperation rates. In experiments, players often cooperate. Without developing a theoretic model of positive reciprocity here (see, e.g., Falk & Fischbacher, 2006), in light of the broad experimental evidence on voluntary public-good games (e.g., Isaac et al., 1985, or the recent surveys by Zelmer, 2003, or Gächter & Herrmann, 2009), we expect players to contribute to the public-good. Furthermore, as shown by Ostrom et al. (1992), Fehr and Gächter (2000), and many others, players are willing to sacrifice own payoff in order to punish others.

In this respect, one has to distinguish between prosocial punishment and antisocial punishment. The literature on public-good games refers to prosocial punishment if the punished player contributes to the public-good less than the norm (e.g., Herrmann et al., 2008). Thus, a norm is a(n implicitly agreed upon) reference value, or contribution target, the deviation from which is deemed inappropriate by the group of interacting players, and therefore leads to deviating players being sanctioned. This sanctioning is referred to as prosocial if it can be interpreted as the attempt to reduce free-riding. In contrast, if the player contributes more than the norm, punishing this player is characterized as being antisocial.[6] Potential explanations for antisocial punishment are, for example, a taste for conformity, revenge, or simply spite.

Despite its importance for human interactions, there is very little evidence concerning the nature of the norms that trigger both prosocial and antisocial punishment. More precisely, there is still some uncertainty about the appropriate reference value to employ when modeling behavior in social-dilemma situations with punishment opportunities. Our study attempts to provide an important empirical step in this respect, contributing to our understanding of sanctioning behavior in public-good games. Thereby, we hope to provide a starting point for future models of norm-related behavior in the broader field of social dilemmas.

When thinking of social norms, a number of questions arises that will be subsequently examined in this article. Carpenter and Matthews (2009) as the only study comparing different norm candidates for prosocial punishment, provide evidence in favor of absolute norms. Notice, however, that this result is obtained in a setting where groups remained constant for the entire duration of the experiment. Thus, one can consider our framework as a robustness check for changing group compositions addressing the question

[6] Others call this form of punishment "perverse", e.g., Cinyabuguma et al. (2006).

RQ 1. *Do absolute contribution norms organize the decisions on whether to announce punishment, to agree to punishment, and how harshly to punish a player better than relative contribution norms?*

Our second research question is concerned with the nature of the norm: does it act only in one direction, explaining punishment of those who underprovide with respect to the norm, or does it also explain punishment of those who deviate positively from the norm? By examining this question, we are able to learn something about the motivation for antisocial punishment. In a post-experimental questionnaire, Fehr and Gächter (2000) asked subjects about the reasons for punishing high-contributors. The answers fall into five categories: (i) random errors; (ii) the contribution level of the high-contributor is still not high enough; (iii) to increase one's relative payoff advantage; (iv) anticipatory revenge against those who might sanction the antisocially punishing player in the current round; and (v) revenge against those who might have sanctioned the player in the previous round (even though, in Fehr and Gächter's case, these could not be identified). In our design, while not impossible, random errors are rather unlikely, as players have to make two random mistakes in a row to exert unwanted punishment: they can always assign 0 points after an announcement.[7] The second category would simply mean that the norm is mis-specified. If this was indeed the case, it would show up in our absolute-norm model as a high absolute norm. Finally, categories (iii)-(v) concern the distinction between point assignements out of revenge, or retaliation, and antisocial punishment not triggered by received punishment points, be it out of spite or competitive thinking. By means of our design, we are able to address this distinction. Therefore, to recapitulate, our second research question is

RQ 2. *Does antisocial punishment – as opposed to retaliation (i.e., punishment triggered by received points) – significantly contribute to explaining decisions on whether to announce punishment and to punish a player? Are there differences over punishment stages?*

Finally, let us discuss the new aspect of our experiment, the elicitation of bystanders' norms of cooperation applied in evaluating others' punishment actions. As described above, we opt to disclose these evaluations publicly, so as not to render them meaningless in the eyes of our subjects. However, the public announcement of others' (dis)agreement may change behavior. Masclet et al. (2003) report a positive effect of (nonmonetary) social

[7]Such errors are rare: in BASIC, the fraction of 0-choices after an announcement is 3%, while it is 16% in OPINION; in the latter, however, the number is largely driven by occasions in which neither player allowed to voice her opinion favored punishment.

(dis)approval on cooperation in public-good games.[8] One reading of this result is that public social assessment of behavior leads to an increase in the degree to which players identify with their group, which in turn may foster cooperation. However, this effect should be much less pronounced – if present at all – in our setting where groups' composition changes after each round. Moreover, in the experiment conducted by Masclet et al., players' voicing of (dis-)approval was an intentional and directed message, rather than a routinely elicited information. Finally, Noussair and Tucker (2007) have shown the effect of social approval to rapidly diminish over the course of the experiment. Hence, whether the display of information on others' evaluations about one's punishment endeavours has any effect on behavior is rather doubtful. A more interesting question is whether players employ different norms when they are in the role of the punisher than when they only act as 'impartial observers'. We therefore set out to answer our final research question, focusing on the relationship between player roles and cooperation norms:

RQ 3. *Are the norm for social approval and the corresponding norms for announcements and punishment identical?*

3 Experimental design

We parameterized our model as follows: let there be $n = 4$ players each endowed with $e = 20$ experimental currency units. We choose $\alpha = 0.4$ and announcement costs equal $f_a = 1$. Finally, for the individual punishment costs, we adopt the cost function used in Fehr and Gächter (2000) and Nikiforakis (2008). The costs for player i punishing player j are given by the convex sequence for increasing $p_{i \to j}$ shown in Table 1.

For recruitment, we use the software package ORSEE (Greiner, 2004), the experimental software is written using z-tree (Fischbacher, 2007); experiments are run at the University of Bonn Experimental Economics Laboratory (BonnEconLab). On the day, subjects are welcomed and asked to draw lots, in order to assign each of them to a cabin. They are asked to move to their cubicle straight away. Once all subjects are seated, the instructions are handed to them in written form before being read aloud by the experimenter.[9] Subjects are given the opportunity to ask any questions concerning

[8] Rege and Telle (2004) come to the same conclusion after conducting a treatment in which they remove players' anonymity altogether. There are interesting variations of public-good games with voting on (non-)enforced absolute cooperation norms (e.g., Walker et al., 2000, Markgreiter et al., 2004, Kroll et al., 2007) and voting on providing or refunding the public-good (Fischer & Nicklisch, 2007).

[9] At the beginning of the experiment, subjects are informed that an unspecified and

Table 1: *Individual punishment costs*

$p_{i\to j}$	0	1	2	3	4	5	6	7	8	9	10
$c(p_{i\to j})$	0	1	2	4	6	9	12	16	20	25	30

the game privately. After questions have been answered individually, subjects are handed a questionnaire to test their understanding of the rules.[10] Questionnaires are corrected individually, while wrong answers are explained privately.

Subjects play ten repetitions (*rounds*) of the game. To prevent the possibility of forming an individual reputation, every player receives an identification number between 1 and 4 at the beginning of each repetition, which she retains for the duration of the round, but which changes randomly in the next one. Furthermore, in order to prevent the emergence of group-specific cooperation norms and to test whether there is a "global" norm for contributions to the public-good, we randomly form groups anew at the beginning of each round out of a pool of 12 subjects ('stranger matching'), while the group composition remains constant within each round.

Altogether, 144 subjects, mostly students majoring in various fields participated in the experiment. Mean age was 24.3 years (standard deviation 6.7 years), 43 percent were females. Each subject participated only once in the experiment. Overall, our data set consists of twelve independent groups of twelve subjects each yielding six independent observations for each treatment condition. Subjects were paid according to the sum of accumulated payoffs gained within the ten repetitions. The experimental currency was converted into Euros and paid off individually to ensure players' anonymity. Each session lasted for approximately 120 minutes, subjects earned on average 18.20 Euros (standard deviation 9.16 Euros, including a 4 Euros show-up fee).

unrelated second part will follow the public good experiment. This second part consists of an unincentivized questionnaire concerning socio-demographic background information of participants.

[10] For a translated version of the instructions and the questionnaire, see Appendices A and B.

4 Results

4.1 Data overview

In Figure 1, we depict round-wise payoffs, contributions, and punishment aggregated over all matching groups for each treatment. Even though contributions start out slightly higher in OPINION (12.9 vs 10.1; contribution levels in the first, second, and third round are different at a level of $p = 0.0782$, $p = 0.1093$, and $p = 0.1495$, respectively), this difference wears away very quickly. In line with the findings of Noussair and Tucker (2007), we do not find any difference in later rounds, nor in the overall contribution level.[11] In the final round, we observe average contributions of half the endowment in both treatments. Furthermore, we do not find any significant differences for aggregate punishment nor efficiency levels as measured by average payoffs. In both treatments, average payoffs start just above the Nash-equilibrium benchmark of 20 experimental currency units and oscillate around a value of 24.5 units towards the end. Average punishment points assigned, on the other hand, fall from 1.2 in the first round to approximately 0.3, in the final two, for both treatments. The average number of punishment iterations is only insignificantly higher in OPINION (1.92 vs 1.72 in BASIC, $p = 0.8095$).[12] Finally, we do not find any significant differences in punishment levels when testing each iteration individually, either.

Looking at the decision of whether to punish or not, we find that overall, about 6% of all possible announcements are made (5.7% in BASIC, 6.2% in OPINION). The time trend mirrors that of punishment in general: whereas in the first round, 8.7% (7.8%) in BASIC (OPINION) of the potential announcements are made, the corresponding figures for the final round are 3.7% for both treatments. Again, the reported treatment differences are far from being significant.[13] On the iterations dimension, we find the highest announcement rate on the first punishment stage (7.2%), followed by the the third and second iterations with 5.3% and 4.1%, respectively.[14]

Drawing on the lack of significant treatment differences in the above, we are confident that our treatment variation does not alter the game substan-

[11] The corresponding values are $p = 0.2002$ for the fifth round, $p > 0.4$ for all remaining rounds, and $p = 0.6991$, for the overall contribution level. Unless otherwise indicated, all (within-)treatment comparisons are done by two-tailed Mann-Whitney U-tests (Wilcoxon signed-rank tests) on the basis of matching-group averages.

[12] This difference is reversed for medians, with medians of 2 in BASIC vs 1 in OPINION.

[13] The corresponding p-values are $p = 0.9372$, $p = 0.6291$, and $p = 0.6171$, for the overall announcement level and the first- and final-round levels, respectively.

[14] In the fourth iteration, we observe a rate of 4.3%, and for the pooled remaining iterations, the figure is 5.1%.

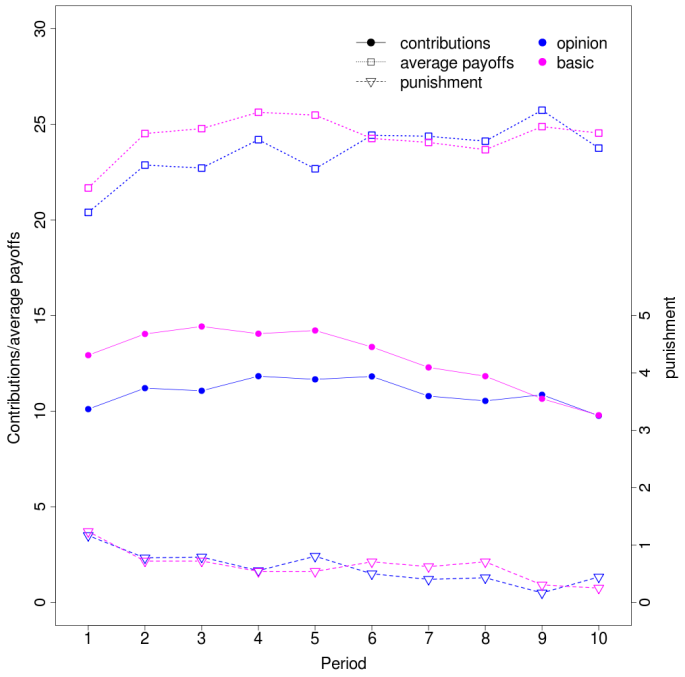

Figure 1: Average payoffs, contributions (both: left axis), and punishment (right axis) over time.

tially.[15] Therefore, we will pool the data from both treatments in our ensuing analysis of contribution norms. The findings from our regressions reported below – namely the non-significance of our treatment dummies – corroborate this claim. We take this as an indication that we can interpret the norms elicited from non-punishers through their endorsement or dissent in the same way as those elicited from the punishers by observing their actions *irrespective* of whether other players are allowed to voice their opinion.

4.2 Contribution norms

To identify the determinants of players' behavior in our public-good game, we will compare the influence of two relative and 21 absolute norms for all three punishment-related decisions of our experiment: the decision to announce punishment, the 'opinion decision', and the actual punishment decision. For each iteration, we will estimate coefficients and absolute norms separately, so that we can identify whether the estimated cooperation norms are stable across iterations. Notice that the number of instances of ongoing iterations

[15] Furthermore, the results are very similar to those reported by Nikiforakis and Engelmann (2009) for their "escalation" treatment.

beyond the third decreases rapidly, so that, in order to rely on a sufficient number of observations, we have to restrict our analysis to the first three iterations of each round.

For the analysis of announcements as well as of the opinions elicited we apply a probit regression with individual error clusters. Thus, we estimate the vector β for the basic econometric models

$$\text{probit}^{-1}(\text{Prob}(a_{i\to j}^{t,m}=1)) = \mathbf{x}'\beta + \varsigma_i + u_{t,m} \ , \tag{3}$$

and

$$\text{probit}^{-1}(\text{Prob}(v_{k:i\to j}^{t,m}=1)) = \mathbf{x}'\beta + \varsigma_k + u_{t,m} \ , \tag{4}$$

where $\text{Prob}(a_{i\to j}^{t,m}=1)$ ($\text{Prob}(v_{k:i\to j}^{t,m}=1)$) stands for the latent probability that i announces to punish j in round t and iteration m (that k endorses i's announcement to punish j in round t and iteration m), $\mathbf{x}$ for the matrix of regressors, β for the vector of coefficients, ς_i for a vector of (unobserved) individual error clusters, and $u_{t,m}$ for a vector of uncorrelated errors.

For the analysis of punishment decisions, we apply a tobit regression with individual error clusters. Thus, for the basic econometric model

$$\hat{p}_{i\to j}^{t,m} = \mathbf{x}'\beta + \varsigma_i + u_{t,m} \ ,$$

and

$$p_{i\to j}^{t,m} = \begin{cases} 10 & \text{if } \hat{p}_{i\to j}^{t,m} > 10, \\ \hat{p}_{i\to j}^{t,m} & \text{if } 0 < \hat{p}_{i\to j}^{t,m} \leq 10, \\ 0 & \text{if } \hat{p}_{i\to j}^{t,m} \leq 0, \end{cases} \tag{5}$$

we estimate the vector β, where $\hat{p}_{i\to j}^{t,m}$ stands for the latent number of punishment points i assigns to j in round t and iteration m, and $p_{i\to j}^{t,m}$ is restricted to the interval $[0, 10]$.

In our quest to identify the norm governing punishment, we compare four models each for the announcement decision, the voiced opinions, and the punishment decision. The first model contains neither an absolute nor a relative norm, but only the control variables (see below), allowing us to assess the importance of either norm for punishment by comparison to the first model. The second and third models test the importance of different relative norms. In the second model, we focus on a group's average contribution. For this purpose, let us define two distance measures. The first, denoted by $r_\star^-$, is the absolute difference between the contribution of the player to be

punished, j, in round t and the average contribution in the group, $\overline{x}^t$, if the former is smaller than the latter, and zero, otherwise:

$$r_{\star}^{-} := |\min\{x_j^t - \overline{x}^t, 0\}|. \tag{6}$$

This variable decreases in the punishee's contribution as long as this contribution is below the group average. A significant positive effect of $r_{\star}^{-}$ would indicate that prosocial punishment is guided by this first relative norm. The second variable, denoted by $r_{\star}^{+}$, reflects positive deviations from the group average and measures the distance between the contribution of player j and the average contribution in the group in round t if the contribution is larger than the average, while the variable is zero otherwise:

$$r_{\star}^{+} := \max\{x_j^t - \overline{x}^t, 0\}. \tag{7}$$

If there is antisocial punishment determined by the first relative norm, we would expect to find a significant positive effect of $r_{\star}^{+}$.

The third model tests the importance of another relative norm, the absolute distance between the contribution of player j and the contribution of the punishing player i in round t. For this purpose, let us define two measures $r_{\star\star}^{-}$ and $r_{\star\star}^{+}$, as

$$\begin{array}{rcll} r_{\star\star}^{-} & := & |\min\{x_j^t - x_i^t, 0\}|, & \text{and} \\ r_{\star\star}^{+} & := & \max\{x_j^t - x_i^t, 0\}. & \end{array} \tag{8}$$

Notice that we retain the reference point of the punisher contribution (x_i^t in equation (8)) in the regressions on voiced opinions, even though it is the bystander taking the decision, so that there could potentially be a change in the reference point. However, a model taking the bystander's contribution as a reference point (not reported here) is clearly outperformed by the reported model 3 on all iterations. A significant positive effect of $r_{\star\star}^{-}$ would indicate that prosocial punishment is guided by the second relative norm, while a significant positive effect of $r_{\star\star}^{+}$ would indicate that antisocial punishment is guided by the second relative norm.

Analogously, the fourth model tests the importance of absolute norms. As in Carpenter and Matthews (2009), we do not allow the absolute norm to change over time in order to increase our ability to distinguish between the absolute and the relative norms. Two regressors measure the absolute distance between the contribution of player j in round t and an integer number $y, y \in [0, 20]$, defined in analogy to the variables measuring the relative norms above:

$$\begin{array}{rcll} a^{-} & := & |\min\{x_j^t - y, 0\}|, & \text{and} \\ a^{+} & := & \max\{x_j^t - y, 0\}. & \end{array} \tag{9}$$

We expect to find a significantly positive effect for the first (second) variable, if there is prosocial (antisocial) punishment that is governed by the absolute norm. Based on a grid search over all possible contribution choices, we select and report that absolute norm fitting the data best according to the log likelihood. This grid search is conducted for each decision and each iteration separately, so that we allow absolute norms to differ. However, assuming that there is an absolute standard guiding behavior, we should observe a consistent y over the different decisions and iterations.

Along with the influence of relative and absolute norms, we control for a number of other regressors that may influence the decisions. For the analysis of the decisions on whether to announce punishment, and of how strongly to punish, those variables include the contribution of the player who punishes (x_i^t) and the sum of contributions of the two players not involved (X_k^t) from that particular round. We expect to find positive effects for both as non-cooperators are typically prosocially punished by players who contribute a substantial amount to the public-good (see, e.g., Cinyabuguma et al., 2006), while free-riders may be more likely to be punished in cooperative groups for reasons of conformity. For potential temporal influences (e.g., learning over the course of the experiment) we test by adding the variable *round*. Moreover, the dummy variable *opinion* marks those decisions from the OPINION treatment. Additionally, for punishment decisions, we also include the variable sum_v^t which counts the number of other players in favor of the punishment action in the OPINION treatment, and which is zero for all observations from the BASIC treatment. Therefore, for punishment points, a negative (positive) effect of *opinion* indicates that there are less (more) points assigned in OPINION than in BASIC if none of the players agrees with the punishment action in the former. However, a negative (positive) effect of sum_v^t indicates that in OPINION, less (more) points are assigned if more of the others consent.

For the analysis of elicited opinions, we have to consider that all observations come from the OPINION treatment (thus, there is no treatment variable in this regression), and that decisions are made by one of the 'third parties'. Therefore, instead of the sum of contribution of the two players not involved, a regressor for the contribution of the player voicing her opinion (x_k^t) is included. Here, similar to the argument that players contributing larger amounts to the public-good are more likely to punish, we expect to find a positive effect of the bystander's own contributions on the endorsement of punishment announcements.

Finally, for the regressions on decisions made in the second (third) iteration, we test for the potential effect of retaliation by means of the variable $p_{j\to i}^{t,1}$ ($p_{j\to i}^{t,2}$) which measures the number of punishment points player i receives from j in the first (second) iteration. This variable – in conjunction

with the term for positive deviations from the norm – allows us to answer our research question **RQ 2**: if punishment of high-contributors is guided by retaliation only, we should see significant effects of $p_{j\to i}^{t,m}$ and no positive effect of a^+, $r_\star^+$, or $r_{\star\star}^+$, respectively. If, however, there is antisocial behavior unrelated to revenge as a motive, the latter variables' coefficients should be significantly different from zero. For $p_{j\to i}^{t,m}$ we expect this to be the case, as according to the findings of Nikiforakis (2008), including a second punishment stage in a public-good game may trigger severe retaliation. In order to analyze differences in retaliation across the two treatments, we include the interaction effect $p_{j\to i}^{t,1} \times \textit{opinion}$ ($p_{j\to i}^{t,2} \times \textit{opinion}$) in our regressions on announcements and on punishment points.

Results for the estimations of mean marginal effects on announcements are reported in Table 2. In all regressions, an absolute term is included, which, however, is not reported. We compare between the nested models (model 1 versus model 2, 3, and 4, respectively) on the basis of the Wald-chi^2-test. Given models 2, 3, and 4 have the same number of regressors, we choose between them by simple likelihood comparisons. Asterisks indicate significance levels.[16]

The results for the first iteration indicate that the probability to announce the punishment of another player increases in the punisher's contribution to the public-good. This holds true even for the third model, although the argument is a little more complex: in this model, we test for the influence of the distance between the punisher's and the punishee's contribution. For that reason, the coefficient for the punisher's contribution x_i^t measures the influence of the level of *both* the punisher's *and* the punishee's contributions *for a given distance*. On the other hand, for a given punishee contribution, an increase in the announcing player's contribution leads to a higher distance $r_{\star\star}^-$, and thus, a higher probability of announcement, as stated above. For three of our models, we also find a significant (positive) effect of increasing contributions of the players being neither the punisher nor the target of the punishment action. Finally, the likelihood of an announcement decreases in the course of the experiment.

With respect to our research question **RQ 1** concerning the norms of contribution in the game, we find that in the first iteration, the relative norm $r_{\star\star}^-$ outperforms the absolute norm in contrast to previous findings by Carpenter and Matthews (2009). Particularly, it seems that players measure others' cooperation levels against the standard set by their own contribution.

[16] *** indicates significance at a $p < 0.01$ level, ** at a $p < 0.05$ level and * at a $p < 0.1$ level. Asterisks attached to log-likelihood values indicate the significance level of the Wald-chi^2-test comparing model 1 and the respective model.

Table 2: *Mean marginal effects for announcements*[16]

iteration		model 1	model 2	model 3	model 4
one	x_i^t	0.006^{***}	0.003^{***}	-0.002^{**}	0.006^{***}
	X_k^t	0.001^{***}	-0.0003	0.003^{***}	0.003^{***}
	round	-0.002^{***}	0.001^{***}	-0.001^{**}	-0.001^{***}
	opinion	-0.012	-0.005	-0.006	-0.003
	$r_\star^-/r_{\star\star}^-/a^-$		0.015^{***}	0.011^{***}	0.009^{***}
	$r_\star^+/r_{\star\star}^+/a^+$		-0.003	-0.002	-0.003
	best absolute norm				15
	log likelihood	-1027.5	-811.5^{***}	-798.1^{***}	-819.3^{***}
two	x_i^t	0.001	0.0004	-0.0003	0.008^{*}
	X_k^t	0.0003	0.0001	0.0005	0.005^{*}
	round	-0.001^{***}	-0.001^{***}	-0.0006^{***}	-0.001^{***}
	opinion	-0.007	-0.009	0.009	0.009
	$p_{j\to i}^{t,1}$	0.015^{***}	0.016^{***}	0.016^{***}	0.016^{***}
	$p_{j\to i}^{t,1} \times$ *opinion*	-0.003	-0.004	-0.004	-0.004
	$r_\star^-/r_{\star\star}^-/a^-$		0.002^{**}	0.002^{***}	0.003^{***}
	$r_\star^+/r_{\star\star}^+/a^+$		-0.001	-0.0005	0.0002
	best absolute norm				10
	log likelihood	-386.1	-377.0^{**}	-373.9^{**}	-372.5^{***}
three	x_i^t	0.0001	0.0002	0.0002	0.0001
	X_k^t	0.001^{**}	0.001^{***}	0.001^{**}	0.001^{**}
	round	0.0003	0.0003	0.0003	0.0003^{*}
	opinion	0.007	0.007	0.007	0.006
	$p_{j\to i}^{t,2}$	0.012^{**}	0.011^{**}	0.012^{**}	0.011^{**}
	$p_{j\to i}^{t,2} \times$ *opinion*	-0.005	-0.004	-0.004	-0.005
	$r_\star^-/r_{\star\star}^-/a^-$		0.0002	-0.0002	0.001^{**}
	$r_\star^+/r_{\star\star}^+/a^+$		0.001	0.00002	0.003^{***}
	best absolute norm				16
	log likelihood	-169.6	-169.2	-169.5	-164.5^{*}

Still, both the $r^-_\star$ norm and an absolute norm of 3/4 of the endowment also indicate the significant influence of a social norm in prosocial punishment behavior, although less accurately so.

Contrary to our first-iteration findings, the results for the second and third iterations favor the absolute norm. For the second, the average-referential norm $r^-_\star$ is clearly outperformed by the self-referential norm $r^-_{\star\star}$ and an absolute norm of 1/2 of the endowment, which does even better. While the effects found in iteration one largely carry over to the second iteration, albeit in a less pronounced manner, retaliation kicks in strongly, as evidenced by the positive effect of first-round punishment $p^{t,1}_{j\to i}$ on the announcement probability. At the same time, the terms for positive deviations from the norm remain insignificant and negative. Thus, answering **RQ 2** for the second iteration, we do not find evidence for motivations for antisocial punishment other than retaliation. The substantial drop in the absolute norm, on the other hand, reflects another effect: players making good for punishment actions not carried out in the first iteration, to ensure the very-low-contributors do not get away without a significant penalty.

For the third iteration, none of the relative norms contributes to explaining subject behavior, while an absolute norm of approximately 3/4 of the endowment does. Furthermore, we observe a significant positive effect of upward deviations from the absolute norm. This can be seen as an indication that, in the third iteration, spiteful or competitive thinking finds its way into the game, giving rise to antisocial punishment. Yet, we cannot exclude the possibility that a different motivation that has not been looked at so far is driving the finding: sanction enforcement, in the sense of punishment of those who fail to punish norm-violators in the first place (cf., e.g., Henrich and Boyd, 2001, for an evolutionary model of cooperation through sanction enforcement). There is an indication that this may actually be a better explanation for our data: there is no inherent need for a spiteful or competitively thinking person to wait before reducing others' income, while for sanction enforcement, the enforcer has to wait for at least one iteration. Interpreting the results in this way, we observe a mix of three different motivations for sanctions: high-contributors lashing back at retaliating low-contributors, low-contributors retaliating against second-iteration punishment after enduring sanctions in iteration one, and high-contributors punishing other high-contributors for not taking part in the sanctioning of low-contributors. The significant positive effect of upward deviations from the absolute norm of 16 seems to suggest that sanction enforcers hold their peers (in contributions) to higher moral standards, the more these had been cooperating. Surprisingly, there is a significant positive effect of *round* in iteration 3, suggesting that arguments tend to become more intense over time. In other words, it is

Table 3: *Mean marginal effects for opinions*[16]

iteration	regressor	model 1	model 2	model 3	model 4
one	x_i^t	0.004^{***}	0.001^{**}	-0.0004	0.002^{***}
	x_k^t	0.002^{***}	0.0003	0.001^{***}	0.001^{***}
	$round$	-0.001^{**}	-0.0002	-0.0001	-0.0002
	$r_\star^-/r_{\star\star}^-/a^-$		0.006^{***}	0.002^{***}	0.004^{***}
	$r_\star^+/r_{\star\star}^+/a^+$		-0.003	-0.0005	-0.002
	best absolute norm				15
	log likelihood	-690.3	-511.2^{***}	-628.6^{***}	-509.3^{***}
two	x_i^t	0.0003	0.0002	-0.0003	0.0004
	x_k^t	0.0003	0.0002	0.0004	0.0004
	$round$	-0.0001	-0.001	-0.0001	-0.0001
	$p_{j\to i}^{t,1}$	0.005^{**}	0.005^{**}	0.004^{**}	0.005^{**}
	$r_\star^-/r_{\star\star}^-/a^-$		0.001	0.001^{***}	0.017
	$r_\star^+/r_{\star\star}^+/a^+$		-0.0004	0.0001	-0.0003
	best absolute norm				1
	log likelihood	-212.4	-209.8^{*}	-208.5^{***}	-208.3^{**}
three	x_i^t	0.0003	0.0002	0.0001	0.0003
	x_k^t	0.001	0.001	0.001^{*}	0.001^{*}
	$round$	0.0003	0.0003	0.0003	0.0003
	$p_{j\to i}^{t,2}$	0.006^{***}	0.005^{***}	0.005^{***}	0.004^{***}
	$r_\star^-/r_{\star\star}^-/a^-$		0.001	0.0003	0.001^{**}
	$r_\star^+/r_{\star\star}^+/a^+$		-0.0003	-0.0001	0.002^{**}
	best absolute norm				15
	log likelihood	-134.3	-132.3	-128.6	-128.3^{***}

less likely that arguments are started in later rounds of the experiment, as evidenced by the decrease in first-iteration announcements. Yet, if they are, they tend to last longer the more rounds have already passed.

Let us now turn to our research question **RQ 3**, asking whether those not directly affected by a punishment action employ the same contribution norms as punishers themselves when evaluating an announcement. An interesting result to be seen from Table 3 is that for explaining players' opinions about others' announcements, the absolute norm performs best on all three iterations. More specifically, the relative norm performing best on punishment actions in iteration one, the punisher's contribution, does poorly in explaining approval in the same iteration, even though it gives the absolute norm a hard race in iterations two and three. This suggests that bystanders tend to evaluate (first-iteration) *prosocial* punishment actions against absolute

standards rather than the adequacy of a relative norm.

The fact that the role of norms in higher iterations is rather limited seems to be owed to the fact that prosocial punishment loses its predominance: when subjects judge retaliative actions, they seem to be guided by the severity of punishment the retaliator has had to endure rather than by contribution levels (cf. the line determined by $p^{t,1}_{j\to i}$ in Table 3). In other words, bystanders tend to find it acceptable that victims of unduely harsh sanctions 'defend' themselves. Correspondingly, it is only in iteration 1 that we find that a punisher will meet stronger endorsement the higher her own cooperation level is, and the more the punishee's contribution falls short of the 3/4 benchmark, where the second effect is about twice as strong. Taking these facts together, this suggests that players are very effective in singling out the motivations of different punishment actions.

What is notable with respect to **RQ 3** is that in iterations one and three, the absolute norm estimated is essentially the same 3/4 of players' endowment we already found for announcements on these iterations. Only for the second iteration do we find a norm differing quite dramatically from that found for punishment announcements. While positive deviations from the absolute norm of 1 do not significantly contribute to explaining endorsements, there is a relatively large increase in agreement if the player to be punished free-rode completely. Given the absolute norm selects complete free-riding to have a differential (yet insignificant) effect, it seems that a sanctioned player's complete uncooperativeness mitigates bystanders' compassion.

Turning to the punishment decisions, we find similar results to the ones for announcements. This also holds with respect to our research questions **RQ 1** and **RQ 2**. As can be seen from Table 4, there is no indication of antisocial punishment that has not been triggered by received points, but strong evidence for retaliation. In the first iteration, the relative norm $r^{-}_{\star\star}$ outperforms the average-referential norm $r^{-}_{\star}$, as well as the absolute norm that is, once again, estimated to be 3/4 of the endowment. In iterations 2 and 3, the absolute norm performs better. Notice further that the absolute norms performing best in predicting the level of punishment points assigned are exactly the same as those estimated in our announcements analysis across all iterations. Thus, although estimated separately, both decisions seem to rely on the same norm of approximately 3/4 of the endowment (unless the norm is confounded by retarded sanctions as it commonly happens in iteration 2).

Similar to what has been said for the norms, our results are similar to those for the announcement decision in all iterations also with respect to the influence of other variables. Additionally, approval of an action has a significant influence on the points assigned across iterations, as evidenced by the significant effect of sum^t_v (i.e., the number of players in favor of the

Table 4: *Mean marginal effects for punishment*[16]

iteration		model 1	model 2	model 3	model 4
one	x_i^t	0.136^{***}	0.084^{***}	-0.078^{**}	0.175^{***}
	X_k^t	0.045^{***}	0.005	0.089^{***}	0.097^{***}
	$round$	-0.057^{**}	-0.032	-0.029	-0.035^{*}
	sum_v^t	5.036^{***}	3.270^{***}	3.216^{***}	3.348^{***}
	$opinion$	-2.760^{***}	-1.996^{***}	-2.031^{***}	-1.994^{***}
	$r_\star^-/r_{\star\star}^-/a^-$		0.434^{***}	0.337^{***}	0.308^{***}
	$r_\star^+/r_{\star\star}^+/a^+$		-0.097	-0.067	-0.079
	best absolute norm				15
	log likelihood	-1220.3	-1107.8^{***}	-1100.2^{***}	-1111.1^{***}
two	x_i^t	0.084^{**}	0.055	-0.018	0.098^{**}
	X_k^t	0.034	0.015	0.058	0.064
	$round$	-0.082^{**}	-0.075^{*}	-0.072^{*}	-0.076^{*}
	sum_v^t	6.119^{***}	5.711^{***}	5.637^{***}	5.696^{***}
	$opinion$	0.416	0.416	0.633	0.703
	$p_{j\to i}^{t,1}$	1.777^{***}	1.977^{***}	1.815^{***}	2.019^{***}
	$p_{j\to i}^{t,1} \times opinion$	-0.536	-0.640	-0.611	-0.693
	$r_\star^-/r_{\star\star}^-/a^-$		0.243^{**}	0.209^{***}	0.368^{***}
	$r_\star^+/r_{\star\star}^+/a^+$		-0.095	-0.015	0.042
	best absolute norm				10
	log likelihood	-480.2	-473.8^{*}	-472.4^{**}	-469.9^{***}
three	x_i^t	-0.021	-0.015	0.020	-0.028
	X_k^t	0.077^{***}	0.081^{***}	0.066^{**}	0.069^{***}
	$round$	0.100^{***}	0.100^{***}	0.101^{***}	0.107^{***}
	sum_v^t	4.197^{***}	4.346^{***}	4.314^{***}	4.205^{***}
	$opinion$	-2.140^{***}	-2.217^{***}	-2.231^{***}	-2.278^{***}
	$p_{j\to i}^{t,2}$	-3.705^{***}	-3.641^{***}	-3.642^{***}	-3.710^{***}
	$p_{j\to i}^{t,2} \times opinion$	4.319^{***}	4.272^{***}	4.295^{***}	4.372^{***}
	$r_\star^-/r_{\star\star}^-/a^-$		-0.122	-0.068	-0.005
	$r_\star^+/r_{\star\star}^+/a^+$		0.010	0.038	0.224
	best absolute norm				16
	log likelihood	-343.6	-341.9	-342.2	-341.3

punishment action in OPINION). If there is no approval, the number of punishment points decreases significantly on iteration one and three, shown by the negative marginal effect of *opinion*, while it does not have an influence on the second iteration. In our interpretation of second-iteration punishment behavior, this corresponds to the substantial amount of retaliative punishment. The authors of such punishment actions may be assumed to factor in the disapproval by society implicitly, announcing only those actions they would carry out irrespective of social approval. On the other hand, the number of points increases substantially if at least one player approves of the punishment action.

Remarkably, there is one result for punishment that does not seem to fit into the broader picture painted so far: on the third iteration, the number of punishment points assigned decreases in the number of punishment points received from the punishee on iteration 2, an effect that is nullified or even reverted by the presence of an 'opinion poll'. What this seems to suggest is that retaliation tends to calm down in iteration 3 in BASIC as indicated by the significant negative marginal effect of $p^{t,2}_{j\to i}$, while the interaction of this variable with the dummy *opinion* suggests that punishment and counter-punishment sequences continue in OPINION. In other words, the public feedback on social approval seems to entrench opposing parties in their positions, so that arguments are fought out more intensely in terms of the punishment level (but not in terms of the number of actions, as our announcement analysis shows).

5 Discussion

In a recent study, Carpenter and Matthews (2009) found that cooperation norms employed in a social-dilemma situation tend to be of an absolute character. In their study, experimental subjects seem to evaluate behavior against an absolute number rather than relative to their own or their group's behavior. This finding is noteworthy, as scholars have mostly restricted their attention to relative measures when attempting to elicit cooperation norms. However, the absolute norms Carpenter and Matthews found for the decision on whether to assign punishment points and that on how many to assign differed substantially from each other, a result that, if robust, would pose a serious challenge to existing theories on the motivations of punishment.

To obtain a better understanding of subjects' cooperation norms, and to dig deeper into how they determine different sanction-related decisions, we extend the line of research pioneered by Carpenter and Matthews with respect to three important dimensions. To disentangle retaliation from punishment related to norms of contribution, we limit interactions to being

one-shot events, having players change their groups in an anonymous and random fashion after each run of the game. By also introducing multiple punishment stages, we achieve three ends: (i) we further separate retaliation from contribution-related sanctioning, as retaliators no longer have to engage in 'pre-emptive counter-punishment'; (ii) we facilitate the distinction of retaliative punishment from antisocial actions driven by other motivations, such as spite or competitive thinking, in our regression analysis; and (iii) we contribute to understanding behavior in a realistic scenario that studies like Denant-Boemont et al. (2007) or Nikiforakis (2008) have shown to lead to substantially different behavior from what is usually observed in public-good experiments with peer punishment as exemplified by Fehr and Gächter (2000). Furthermore, to obtain a clearer picture about whether the decisions to punish and how many points to assign are driven by different processes, we explicitly have our subjects take these decisions separately. Finally, we introduce a second treatment to provide us with data on how bystanders evaluate punishment actions, an information that, to the best of our knowledge, has not been looked at by any preceeding studies.

Our findings are noteworthy in a number of ways. First of all, we find support for a finding already made by Carpenter and Matthews: the average-related contribution norm $r_{\star}^{+/-}$, estimated in a non-negligible number of important contributions like, e.g., Fehr and Gächter (2000, 2002) or Anderson and Putterman (2006) is outperformed as a predictor of behavior by other models on every iteration and each decision. In other words, our data provides evidence for the reasons behind the shift towards the punisher's own contribution as the norm to be estimated in recent studies such as, e.g., Herrmann et al. (2008) or Egas and Riedl (2008).

Furthermore, like Carpenter and Matthews, we find strong support for the influence of an absolute cooperation norm. However, on the first iteration, in both decisions taken by the punisher this norm is outperformed by the relative norm set by the punisher's own contribution. In contrast, for bystanders' decisions, the absolute norm leads to a better fit across all iterations. This might not be as surprising as it may seem: if we interpret behavior on the first punishment stage as an intuitive reaction to others' cooperation levels, we may indeed expect that behavior to be self-referential. On later stages, in contrast, evaluations of others' choices will be comparatively more detached, and thus, more focused on the absolute level of 'sanction-deservingness' of the punishee – as will the evaluations of others' punishment endeavors. Summing up, the answer to our research question **RQ 1** is to be contingent on the decision concerned: absolute contribution norms organize those decisions that cannot be seen as intuitive first reactions to others contribution decisions better than relative contribution norms; punishment decisions on the

first iteration are best predicted by the self-referential relative norm $r_{\star\star}^{+/-}$.

A third important result is that, in contrast to the finding of Carpenter and Matthews, the best-performing absolute norms are very consistent across different decisions within the different iterations. Not only that, the cooperation norm of 3/4 of players' endowment from the first iteration even carries over to the third one, suggesting a certain stability over time. Remarkably, however, this norm seems to disappear in the second iteration. This is a clear sign that the processes generating our data differ across different iterations.

In the first stage, prosocial punishment by high-contributing players is the predominant factor. Our design keeps this iteration clear from revenge-related point assignments and minimizes assignments due to random errors. On the other hand, our analysis shows that antisocial punishment for other reasons does not play a role, either. In other words, the data we obtain from the first iteration is particularly well-suited for comparing potential candidate variables to build a theory of norm-related punishment on.

In the second iteration, we observe a mix of counterpunishment and retarded sanctions left out in the first stage, but we still do not find any evidence for antisocial punishment for reasons other than revenge. Retarded sanctions, however, will tend to bring the absolute norm down, which is what we observe in our analysis, while retaliators will tend to be relatively insensitive to the contribution levels of their opponents. As a consequence, a lower norm will provide a better fit. An observation that *prima facie* looks surprising is the extremely low absolute norm in bystanders' evaluation in the second iteration. Only players not contributing at all meet less endorsement when announcing punishment on this iteration. At the same time, the only significant determinant of bystanders' agreement with the assignment of points in the second iteration is the number of points priorly received by the announcing player. In particular, subjects agree with retaliation more often the higher the initial sanctions are. In other words, subjects seem to endorse retaliation when sanctions are unduely harsh. The fact that the absolute norm selects the very lower end of the contribution spectrum to have a differential effect, even though not significantly so, seems to suggest that bystanders' compassion towards harshly sanctioned players is mitigated (only) if these players have proven to be completely uncooperative.

In the third iteration, finally, there is another type of punishment intermingling with retaliative action that could be attributed to either spitefulness, competitive thinking, or sanction enforcement, the latter being a type of sanctioning behavior often assumed to be the stabilizing force behind prosocial punishment (e.g., Henrich & Boyd, 2001). Judging by the fact that this type of high-contributor sanctioning shows only late in the interaction, we

tend to favor the latter explanation. In terms of our research question **RQ 2**, the above leads us to conclude that 'unsolicited' antisocial punishment does not play a role in early stages when controling for retaliative actions. Whether this changes in later stages, or whether the punishment of high-contributors we observe in the third iteration is due to sanction enforcement is an important question that will need further research.

As far as our research question **RQ 3** is concerned, we find astonishingly consistent estimates for absolute norms over different decisions within an iteration, and thus, even over different player roles. However, we have to introduce two caveats: (i) in the first iteration, the fact that the absolute norms estimated from both punisher and bystander choices are identical seems to be a coincidence of different processes leading to similar results: in our interpretation, first-stage punishers intuitively react in a self-referential manner after having made a contribution sufficiently close to 3/4 of their endowment, while bystanders take a more detached view, evaluating punishee behavior against a more neutral absolute standard that is also equal to three quarters; (ii) especially in the second iteration, support is primarily driven by what looks like empathy with overly-sanctioned retaliators. Even if the first caveat appears to be a strong point against our hypothesis, we would like to emphasize that the two processes may be closely related: first-stage punishers may view themselves as being entitled to sanction others *because* they complied with the absolute 3/4-norm. Taken together, three quarters of players' endowment seems to be the socially accepted reference point for punisher contributions in our experiment, as well as the standard against which punishees' behavior is evaluated. However, players do distinguish between punishment related to norm violations with respect to (i) contributions and (ii) adequate punishment severity. It is only with respect to the former that our estimated absolute norm provides a robust reference point.

Overall, our experimental results underline the importance of norms for behavior even in a setting with anonymous, self-contained episodes of interaction and changing partners between those episodes. The fact that the estimated norms tend to be consistent over decisions and, to some degree, even over iterations, suggests that we are observing truly *social* norms in our experiment, in the sense that players seem to bring an intuitive understanding of adequate behavior into the laboratory that is likely to be shaped by cultural values rather than being a mere experimental artifact. In this light, we are confident that our results contribute to the understanding of norm-related behavior, enhancing the way economists think about and model this important element of human interaction.

References

[1] Anderson, Christopher M., and Louis Putterman (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.

[2] Carpenter, Jeffrey P., and Peter H. Matthews (2009): What norms trigger punishment? *Experimental Economics* 12, 272-288.

[3] Cinyabuguma, Matthias, Talbot Page, and Louis Putterman (2006): On perverse and second-order punishment in public goods experiments with decentralized sanctioning. *Experimental Economics* 9, 265-279.

[4] Denant-Boemont, Laurent, David Masclet, and Charles Noussair (2007): Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33, 145-167.

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

[6] Falk, Armin, and Urs Fischbacher (2006): A theory of reciprocity. *Games and Economic Behavior* 54, 293-315.

[7] Fehr, Ernst, and Simon Gächter (2000): Cooperation and punishment in public goods experiments. *American Economic Review* 90, 980-994.

[8] Fehr, Ernst, and Simon Gächter (2002): Altruistic punishment in humans. *Nature* 415, 137-150.

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

[10] Fischer, Sven, and Andreas Nicklisch (2007): Ex interim voting: An experimental study of referendums for public good provision. *Journal of Institutional and Theoretical Economics* 163, 56-74.

[11] Gächter, Simon, and Benedikt Herrmann (2009): Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 791-806.

[12] Greiner, Ben (2004): An online recruitment system for economic experiments. In: *Forschung und wissenschaftliches Rechnen 2003: GWDG Bericht* 63, edited by K. Kremer and V. Macho, 79-93. Göttingen: Gesellschaft für Wissenschaftliche Datenverarbeitung.

[13] Henrich, Joseph, and Robert Boyd (2001): Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208, 79-89.

[14] Herrmann, Benedikt, Christian Thöni, and Simon Gächter (2008): Antisocial punishment across societies. *Science* 319, 1362-1367.

[15] Isaac, R. Mark, Kenneth F. McCue, and Charles R. Plott (1985): Public good provision in an experimental environment. *Journal of Public Economics* 26, 51-74.

[16] Kroll, Stefan, Todd L. Cherry, and Jason F. Shogren (2007): Voting, punishment, and public goods. *Economic Inquiry* 45, 557-570.

[17] Masclet, David, Charles N. Noussair, Steven Tucker, and Marie-Claire Villeval (2003): Monetary and nonmonetary punishment in the voluntary contributions mechanisms. *American Economic Review* 93, 366-380.

[18] Margreiter, Magdalena, Matthias Sutter, and Dennis Dittrich (2005): Individual and collective choice and voting in common pool resource problems with heterogeneous actors. *Environmental and Resource Economics* 32, 241-271.

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

[20] Nikiforakis, Nikos, and Dirk Engelmann (2009): Feuds in the laboratory? A social dilemma experiment. *Working Paper.*

[21] Noussair, Charles, and Steven Tucker (2007): Public observability of decisions and voluntary contributions in a multiperiod context. *Public Finance Review* 35, 176-198.

[22] Ostrom, Elinor, James M. Walker, and Roy Gardner (1992): Covenants with and without a sword: Self-governance is possible. *The American Political Science Review* 86, 404-417.

[23] Ostrom, Elinor (2000): Collective actions and the evolution of social norms. *The Journal of Economic Perspectives* 14, 137-158.

[24] Rege, Mari, and Kjetil Telle (2004): The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics* 88, 1625-1644.

[25] Reuben, Ernesto, and Arno Riedl (2009): Enforcement of contribution norms in public good games with heterogeneous populations. *Working Paper.*

[26] Sefton, Martin, Robert Shupp, and James Walker (2007): The effect of rewards and sanctions in provision of public goods. *Economic Inquiry* 45, 671-690.

[27] Sethi, Rajiv (1996): Evolutionary stability and social norms. *Journal of Economic Behavior and Organization* 29, 113-140.

[28] Sober, Elliott, and David S. Wilson (1998): *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Cambridge, MA: Harvard University Press.

[29] Sugden, Robert (1986): *The Economics of Rights, Co-operation, and Welfare.* Oxford: Blackwell Publishing Limited.

[30] Sutter, Matthias, Stefan Haigner, and Martin G. Kocher (2008): Choosing the stick or the carrot? Endogenous institutional choice in social dilemma situations. *Working Paper.*

[31] Walker, James, Roy Gardner, Andrew Herr, and Elinor Ostrom (2000): Collective choice in the commons: Experimental results on proposed allocation rules and votes. *Economic Journal* 110, 212-234.

[32] Yamagishi, Toshio (1986): The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology Review* 51, 110-116.

[33] Zelmer, Jennifer (2003): Linear public goods experiments: A meta-analysis. *Experimental Economics* 6, 299-310.

Appendix A: Instructions[17]

Thank you very much for your participation in this experiment. You are now participating in an economic experiment. If you carefully read the following explanations, you can earn a substantial amount of money, contingent on your decisions. Therefore, it is very important that you read these explanations carefully.

The instructions handed out to you are for your private information only. During the experiment there is a strict prohibition of any kind of communication. If you have any question, please, direct them towards us. If you do not abide by this rule, you will be excluded from the experiment as well as any payments.

During the experiment we will not talk about Euros but about Ecu. Your total payoff will first be calculated in Ecu. The total amount of Ecu you obtain during the experiment will be converted to Euros at the end of the experiment, with 25 Ecu = 1 Euro. At the beginning (and additional to the 4 Euros for showing up), each participants will be given a one-time flat-fee payment of 25 Ecu. Using these 25 Ecu, you may cover potential losses. You can always avoid losses with certainty by making decisions accordingly. You will be paid your earnings in Ecu (including the one-time flat-fee payment) plus 4 Euros for showing up. This will be done privately and in cash.

The experiment will consist of two parts. In the following, the course of part one will be described. The explanations regarding the second part will be given to you later. Altogether, the first part consists of 10 periods. In every period, the experiment will consist of 4 steps. Participants are divided into groups of four. Therefore, apart from yourself your group will contain three other members. However, you do not know the identity of the other participants. In every period, the composition of the group will be newly determined by chance.

The first step

At the beginning of each period, every participant will be provided with 20 Ecu which we will call endowment in the following. Your task is to make a decision on the use of your endowment. You have to decide how many out of the 20 Ecu you deposit into a project (0 to 20) and how many you keep for yourself. The consequences of this decision will be explained in more detail below.

[17] The following instructions are translations of the German originals that were adapted from Nikiforakis (2008) and are available from the authors upon request. Treatment variations are indicated by brackets.

Once all members of the group have decided on their deposits into the project, you are informed about the contributions of the group members, your payoff from the project, and your payoff from step 1. Your payoff is calculated according to the following simple formula:

Your payoff from the first step equals:
20 - (your deposit into the common project) +
0,4 x (sum of deposits of all group members into the common project)

As you see, your payoff from step 1 of a period is composed of two parts:

- Ecu you keep for yourself = endowment - your deposit into the project
- The payoff from the project = 0,4 x sum of deposits of all group members

The payoff from the project of all other group members is calculated using the same formula, i.e., each group member receives the same payoff from the project. If, for example, the sum of deposits of all group members equals 60 Ecu, you and all other group members obtain a payoff of 0.4x60 = 24 Ecu from the project. If the group members deposit a total of 9 Ecu into the project, you and all other group members receive a payoff of 0.4x9 = 3.6 Ecu from the project.

Every Ecu you keep earns you a payoff of 1 Ecu. If, instead, you deposit one Ecu out of your endowment into the project of your group, the sum of deposits will rise by 1 Ecu and your payoff from the project will rise by 0.4x1 = 0.4 Ecu. However, the payoff of all other group members will also rise by 0.4 Ecu, such that the total earnings of the group increase by 0.4x4 = 1.6 Ecu. Therefore, through your deposits into the project, all other group members will also gain something. Conversely, you will also gain something from the deposits into the project of other group members. For each Ecu another group member deposits into the project, you earn 0.4 Ecu.

The second step

In the second step, you are informed about the deposits of the other group members into the project. After that, each group member may announce to assign points to one or several other group members. Each announcement costs you 1 Ecu. Other group members can also announce to assign points to you.

In the third step, you can only assign points to group members you designated on the second step. All group members will be informed about all announcements of point assignments.

[OPINION The two group members not affected by an announcement can approve or reject it. An announcement that has not been approved by at least one unaffected player is considered to be rejected. All group members are subsequently informed about the individual approvals or rejections.]

The third step

In the third step, [OPINION you are informed about the results of all votes in detail. Afterwards,] you determine the level of points. [OPINION The assignment of points can be effected independently of the voting result.] By an assignment of points, the payoff of the corresponding group member is decreased. Other group members can also decrease your payoff if they want. If you choose 0 points for a certain group member, you do not change that group member's payoff. If, however, you assign one point to a member, you decrease the corresponding group member's payoff in Ecu from the first step by 10 percent. If you assign 2 points to a group member, you decrease that person's payoff by 20 percent, etc. In other words, the points you assign determine how much a group member's payoff in Ecu from the first step is decreased. If a person receives a total of 4 points, then that person's payoff from the first step is curtailed by 40 percent. In case a person receives exactly 10 or more points, then that person's payoff from the first step will be reduced by 100 percent.

If you assign points, you incur costs in Ecu that depend on your assignment of points. You may assign between 0 and 10 points to every group member. The more points you assign to a group member, the higher your costs are. The total costs in Ecu are calculated as the sum of costs of points assigned to all other group members. The following table specifies the relationship of assigned points and the costs of assigning points in Ecu:

Points	0	1	2	3	4	5	6	7	8	9	10
Costs of points	0	1	2	4	6	9	12	16	20	25	30

If, for example, you assign 2 points to a member of your group, you incur costs of 2 Ecu; if you additionally assign 8 points to another member, you incur costs of 20 Ecu. Your total costs therefore amount to 22 Ecu (2+20), not 30 Ecu. Additionally, you have to bear costs of 2 Ecu for the announcements.

Your total costs for points, that is, the sum of costs for points assigned to other group members and the sum of costs for announcements will be deducted from your payoff from the first step. Your period payoff after the third step is therefore given by the following formula:

Your period payoff therefore amounts to:
(Your payoff from the first step)(1 - (sum of points you receive)/10)
- (sum of costs for points you assigned) - (sum of costs for announcements)

If you receive more than 10 points from other group members, the maximum amount deducted from you will be your total payoff from the first step. In other words, your payoff from the first step can only be reduced to 0. However, you still have to bear the total costs of points you assigned. Therefore, your period payoff can become negative through according decisions. You can make up for negative period payoffs through the flat-fee payment of 25 Ecu you received at the beginning.

The fourth step

After all participants have made their decisions, they are informed about the points assigned to themselves and about their origin.

If at least one group member has announced the assignment of points on the second step, each group member is, again, allowed to announce the assignment of points to one or several other group members (otherwise the period payoff equals the payoff from the first step and there are no further announcements). Each new announcement again causes a cost of 1 point.

[OPINION Again, those group members not involved may voice their approval.] Afterwards, the level of points may be increased or new points may be assigned.

Please note: if you assign points to a group member you have already apportioned points to within this period, what is relevant for both your period payoff and the affected group member's payoff is the total sum of points, not the sum of the individual assignments. In other words, points assigned to the same group member are added: if, for example, you first assign 2 points and later on another 3 points to a group member, you have to bear total costs of 9 Ecu (and not 2+4 = 6 Ecu), plus 2 Ecu for the announcements.

You can only make announcements or assign points if this does not lower your period payoff below zero. Again, all group members are informed about their current period payoffs and new announcements and assignments of points are possible. This repetition only ends when no group member announces the assignment of further points. If no group member announces the assignment of further points, a new period starts in a newly and randomly composed group.

Total payoff

The total payoff is given by the sum of period payoffs from all periods.

Appendix B: Questionnaire

Please answer all questions. There are no consequences for you due to wrong answers. If you have any questions please contact us.

1. Each group member is endowed with 20 Ecu. None (including you) contributes anything in the first stage.
    - What is your income in the first stage?
    - What is the income of each of the other group members in the first stage?
2. Each group member is endowed with 20 Ecu. Each group member (including you) contributes 20 Ecu to the project in the first stage.
    - What is your income in the first stage?
    - What is the income of each of the other group members in the first stage?
3. Each group member is endowed with 20 Ecu. The other three group members contribute in total 30 Ecu to the project in the first stage.
    - What is your income in the first stage if you contribute – in addition to the 30 Ecu – 0 Ecu to the project?
    - What is your income in the first stage if you contribute – in addition to the 30 Ecu – 15 Ecu to the project?
4. Each group member is endowed with 20 Ecu. You contribute 8 Ecu to the project.
    - What is your income in the first stage if the others group members contribute – in addition to your 8 Ecu – in total 7 Ecu to the project?
    - What is your income in the first stage if the others group members contribute – in addition to your 8 Ecu – in total 22 Ecu to the project?
5. In the second stage you announce to distribute points to each of the three other group members. You distribute 9, 5, and 0 points.

- What are the total costs for the distribution of those points?

6. What are the total costs if you announce to distribute points to one of the group members and distribute 0 points?

7. What is the reduction of first stage income if you receive in total

   - 0 points
   - 4 points
   - 15 points

   from the other group members?

8. You announce to distribute points to two of the three other group members. You distribute 2, and 2 points. Then you announce to distribute points to all three other group members and distribute 1, 1, and 1 point.

   - What are the total costs for the distribution of those points?