Cooperation Norms in Multiple-Stage Punishment

ANDREAS NICKLISCH
University of Hamburg

IRENAEUS WOLFF
University of Konstanz

Abstract

We analyze the interplay between cooperation norms and people's punishment behavior in a social-dilemma game with multiple punishment stages. By combining multiple punishment stages with self-contained episodes of interaction, we are able to disentangle the effects of retaliation and norm-related punishment. An additional treatment provides information on the norms bystanders use in judging punishment actions. Partly confirming previous findings, punishment behavior and bystanders' opinions are guided by an absolute norm. This norm is consistent over decisions and punishment stages and requires full contributions. In the first punishment stage, our results suggest a higher personal involvement of punishers, leading to a nonlinearity defined by the punishers' contribution. In later punishment stages, the personal-involvement effect vanishes and retaliation kicks in. Bystanders generally apply the same criteria as punishers in all stages.

Andreas Nicklisch, University of Hamburg, Germany, and Max Planck Institute, Bonn, Germany (nicklisch@coll.mpg.de). Irenaeus Wolff, University of Konstanz, Germany, and Thurgauer Wirtschaftsinstitut (TWI), Kreuzlingen, Switzerland (wolff@twi-kreuzlingen.ch).

We are deeply indebted to Sophie Bade, Katharine Bendrick, Christoph Engel, Michael Kurschilgen, Bettina Rockenbach, Marie-Claire Villeval, and two anonymous referees 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.

Received September 30, 2009; Accepted September 16, 2010.

Journal of Public Economic Theory, 13 (5), 2011, pp. 791–827.

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 and 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 misaligned. Norm violations and others' responses to such violations have long been debated in the experimental literature in the context of decentralized sanctioning mechanisms. In this context, a norm is the (implicitly agreed upon) reference value of the cooperation level such that deviating from this cooperation target leads to the deviating players being sanctioned.[2]

Sanctions have been shown to foster and maintain voluntary cooperation in social dilemmas (seminal work has been provided by Ostrom et al. 1992, for common-pool resources, and Yamagishi 1986, or Fehr and Gächter 2000, for public goods). Our 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 tailored for 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.[3]

When thinking about cooperation norms in social-dilemma situations, one important distinction is that between relative and absolute norms. Relative norms are variable reference points that rise and drop with the level of cooperation within the 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). A relative-norm model would merely predict punishment to be observed until behavior has converged; an absolute-norm model also specifies the point of convergence.

---

[1] Cf. Ostrom (2000).

[2] Cf. the use of the term, e.g., by Carpenter and Matthews (2009).

[3] Note that we do not analyze how punished players react to sanctions that are justified according to the different norms. Evaluating reactions in this sense would be an interesting exercise, but would require that we assume the crucial norm in advance. Other authors have explored this interesting issue (e.g., Cinyabuguma et al. 2006, Ones and Putterman 2007) which would go beyond the scope of our experimental design.

Relative norms have been estimated in a number of studies, as theoretic models of prosocial behavior like the Fehr and Schmidt (1999) model suggest reference points to be relative in the above sense. This idea has received empirical support by studies such as Dawes et al. (2007) or Johnson et al. (2009) who find evidence for egalitarian motives as a driving factor in costly punishment. In terms of norm choice, several authors rely on the average degree of cooperation within the group as the norm (Fehr and Gächter 2000, 2002, Anderson and Putterman 2006, Sefton, Shupp, and Walker 2007), whereas more recent studies focus on the degree of cooperation of the player who punishes (Herrmann, Thöni, and Gächter 2008, Egas and Riedl 2008, Reuben and Riedl 2009, Sutter, Haigner, and Kocher 2010). 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 is 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. This finding, if robust, would challenge theoretical attempts to explain punishment behavior by existing models of pro-social behavior.

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 counter-punishment 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, Masclet, and Noussair 2007, Nikiforakis 2008, and Nikiforakis and Engelmann 2011).

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. This behavior is categorized as "antisocial punishment" (e.g., Herrmann, Thöni, and Gächter 2008).[4] Cinyabuguma, Page, and Putterman (2006) present some evidence that most antisocial punishment seems to stem from a sort of "blind revenge." Thanks to our design, we are able to draw an even clearer picture and provide evidence on the social acceptability of retaliation. At the same time, we can largely rule out random errors as another possible source of

[4] Others call this form of punishment "perverse," e.g., Cinyabuguma, Page, and Putterman (2006).

high-contributor punishment suggested in the literature (cf. Fehr and Gächter, 2000).

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.[5] Explicitly disentangling the decisions of whether to punish a player and by how much will be interesting, because it allows us to analyze the degree of consistency between the norms.

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 are 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. 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, whereas they provide additional information on norms within a group. Further details concerning the experimental design are discussed in the following two sections.

Our results indicate that in line with the findings of Carpenter and Matthews, an absolute norm seems to organize the decisions relating to norm violations very well. Particularly, we observe an absolute norm defined by subjects' endowments that is consistent over different decisions and different actors. Often, a player's own contribution relative to the punished-to-be's contribution acts as an additional trigger in the first iteration. However, this phases out quickly, as do contributions as a determinant of punishment-related decisions in general, but at a slower pace. In our treatment variation, bystanders' opinions rather than contribution differences serve as the main

[5] Similarly, Masclet et al. (2009) employ a two-step procedure for punishment; in their case, punishment actions are publicly announced *before* the cooperation stage for each possible cooperation level. Subsequently, the announcer can revise her schedule in the actual punishment stage.

determinant of the punishment level. However, opinions follow patterns that are remarkably similar to those found in punishers' announcements, which do not exhibit significant differences between treatments. Due to this fact, the observed behavior in both treatments is hardly distinguishable.

We observe punishment of high-contributors by lower-contributors predominantly as a response to prior sanctioning by the former. This suggests that the "perverse" punishment observed in earlier studies is a form of "blind revenge" or "pre-emptive counter-punishment" rather than spiteful or competitive behavior or the consequence of a taste for conformity; only in our treatment variation, there are instances of "perverse" punishment. However, additional research is needed to clearly determine the reasons for this surprising treatment difference.

The remaining paper is organized as follows: Section 2 introduces the game and presents our research questions. Section 3 describes the experimental design. Section 4 reports the results, whereas section 5 discusses the findings along with their implications.

2. The Game and Research Questions

2.1. The Game

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.[6]

In *step three*, the announcements are made public knowledge, and in our OPINION condition, the players who are neither the punisher nor the target of an announcement $a_{i \to j}$, that is, all players k s.t. $k \notin \{i, j\}$, may voice their

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

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 10%, and therefore, we have a natural limit for punishment points, $p^{\max} = 10$.[7] Therefore, the payoff equals

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

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 2–4 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, for example, 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.[8] At the time of making their punishment-related decisions, players are provided with information about the accumulated points assigned to themselves and about their origin, the accumulated points received by other group members and the resulting payoffs, alongside the initial contributions to the public good made by each of the players. Thus in every iteration, information is provided that is may provide a basis for norm-guided or retaliative punishment. We repeatedly allow for new announcements and increases in punishment points until no player makes a further announcement to punish.[9] Notice that

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

[8] 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.

[9] This procedure is similar to the one used by Nikiforakis and Engelmann (2011) in their multiple-stage treatments.

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.

2.2. Predictions

Because 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 according to standard theory in which any player will only be concerned with his own monetary payoff. On the equilibrium path of the unique subgame-perfect Nash 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 player will contribute to the public-good, because 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 and Fischbacher 2006), in light of the broad experimental evidence on voluntary public-good games (e.g., Isaac, McCue, and Plott 1985, or the recent surveys by Zelmer 2003 or Gächter and Herrmann 2009), we expect players to contribute to the public-good. Furthermore, as shown by Ostrom, Walker, and Gardner (1992), Fehr and Gächter (2000), and many others, players are willing to sacrifice own payoff to punish others.

2.3. Research Questions

When thinking of social norms, a number of questions arise that will be subsequently examined in this paper. In the only study comparing different norm candidates for prosocial punishment, Carpenter and Matthews (2009) 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 following question:

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 un-

derprovide 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 postexperimental 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, although 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.[10] Category (ii) 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 assignments 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 punishment 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 earlier, 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 (dis)approval on cooperation in public-good games.[11] 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—as (i) in our setting, players' voiced (dis-)approval was a routinely elicited information rather than an intentional and directed message

[10] Such errors are rare: in BASIC, the fraction of 0-choices after an announcement is 3%, whereas 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.

[11] 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, Margreiter, Sutter, and Dittrich 2004, Kroll, Cherry, and Shogren 2007) and voting on providing or refunding the public-good (Fischer and Nicklisch 2007).

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

as in Masclet et al. (2003), and (ii) 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 of one's punishment endeavors has any direct effect on contribution behavior is rather doubtful, although it may influence the level of point assignments. Nonetheless, we expect this effect to be rather weak. A more interesting question in terms of our main topic 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: *Does the norm for social approval differ from the norms for both announcements and punishment?*

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 used the software package ORSEE (Greiner 2004), the experimental software was written using z-tree (Fischbacher 2007); experiments were run at the University of Bonn Experimental Economics Laboratory (BonnEconLab). On the day, subjects were welcomed and asked to draw lots, to assign each of them to a cabin. They were asked to move to their cubicle straight away. Once all subjects were seated, the instructions were handed to them in written form before being read aloud by the experimenter.[12] Subjects were given the opportunity to ask any questions concerning the game privately. After questions had been answered individually, subjects were handed a questionnaire to test their understanding of the rules.[13] Questionnaires were corrected individually, although wrong answers were explained privately.

[12] At the beginning of the experiment, subjects were informed that an unspecified and unrelated second part would follow the public good experiment. This second part consisted of an unincentivized questionnaire concerning socio-demographic background information of participants.

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

Subjects played 10 repetitions (*rounds*) of the game. To prevent the possibility of forming an individual reputation, every player received an identification number between 1 and 4 at the beginning of each repetition, which she retained for the duration of the round, but which changed randomly in the next one. Furthermore, 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 formed groups anew at the beginning of each round out of a pool of 12 subjects ("stranger matching"), whereas the group composition remained 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% were females. Each subject participated only once in the experiment. Overall, our data set consists of 12 independent groups of 12 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 (at a rate of 25 units per euro) and subjects were paid 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-euro show-up fee).

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 $p = 0.0782$, 0.1093, and 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.[14] 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 or 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. As such, our results are in line with those of Nikiforakis and Engelmann (2011) whose design is closest to ours, but contrast with the findings by Nikiforakis (2008) and Denant-Boemont, Masclet,

[14] 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.

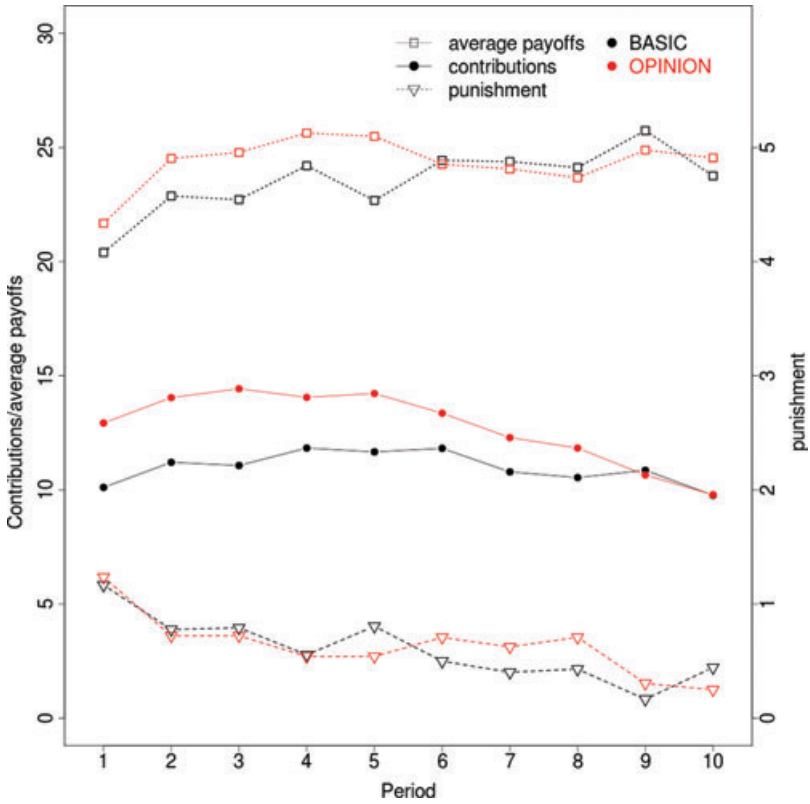

Figure 1: Average payoffs, contributions (both: left axis), and punishment (right axis) over time.

and Noussair (2007). The latter two papers report that in the presence of counter-punishment opportunities, cooperation breaks down. We will discuss potential reasons for the observed difference in the discussion of our paper.

Average punishment points assigned over all iterations of a round 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, $p = 0.8095$).[15]

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

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

both treatments. Again, the reported treatment differences are far from being significant.[16] On the iterations dimension, we find the highest announcement rate in the first punishment stage (7.2%), followed by the third and second iterations with 5.3% and 4.1%, respectively.[17]

Before we proceed to estimate the norms guiding our subjects' punishment behavior, let us take a closer look at the general punishment patterns in the two treatments. For this purpose, we classify punishment actions according to the punishing and punished players' contribution ranks.

4.2. Punishment Patterns

When aggregating the data over all iterations, we note that there is no general treatment difference with respect to the ranks of punishers and punished players; this applies to both announcements and punishment received. To describe punishment behavior in greater detail, we disaggregate the data by iterations. Notice that the number of instances of ongoing iterations beyond the third decreases rapidly, so that to rely on a sufficient number of observations, we have to restrict our analysis to the first three iterations of each round. We find that the frequency of announcements is the same across treatments in iterations 1 and 2, but this frequency has a tendency to be higher in OPINION in iteration 3 ($p = 0.0910$). To analyze punishment patterns further, we test which contribution ranks mete out punishment, and who receives the punishment points. To this end, contributions within the group of four players are ranked: the player with the highest contribution is denoted by "max," the second-highest by "3," and so on.[18] For this exercise, we abstract from the number of points assigned but only count punishment actions. We will elaborate more on the number of points assigned in section 4.3 when discussing the estimated norms.

For a first rough picture of the emerging punishment patterns, we provide Table 2. In this table, we show the frequency of punishment actions by iteration, treatment, and contributor rank, relative to the corresponding punishment opportunities.[19] Looking at the rank of players who are subject to punishment (i.e., comparing columns), there is no significant treatment difference in any of the iterations. On a more general level, by looking at each iteration's lower-left-hand corners in the table the impression may arise

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

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

[18] In case of a tie, contributors are assigned the higher rank (i.e., if there are two players who contributed the second-highest contribution, they both are grouped to "3"). Accordingly, two players tying on the group's smallest contribution are assigned rank 2.

[19] Note that the data provided in Table 2 is an aggregation of all data points, irrespective of their (in-)dependence. Of course, the significance tests following below are conducted based on independent observations.

Table 2: Punishment actions by contribution ranks, as fractions of opportunities

	Punishment in BASIC					Punishment in OPINION					
	max	**3**	**2**	**min**	**overall**	**max**	**3**	**2**	**min**	**overall**	
					ITERATION 1						
max	0.00	0.02	0.10	0.32	0.12	0.03	0.01	0.05	0.24	0.09	max
3	0.00	0.01	0.07	0.28	0.09	0.01	0.00	0.03	0.16	0.05	3
2	0.01	0.00	0.00	0.24	0.05	0.01	0.01	0.00	0.11	0.03	2
min	0.00	0.00	0.01	–	0.00	0.04	0.03	0.04	–	0.04	min
					ITERATION 2						
max	0.01	0.01	0.03	0.06	0.03	0.05	0.03	0.02	0.07	0.05	max
3	0.03	0.00	0.02	0.05	0.03	0.02	0.06	0.04	0.03	0.03	3
2	0.01	0.02	0.04	0.04	0.02	0.02	0.02	0.09	0.06	0.03	2
min	0.04	0.05	0.04	–	0.05	0.07	0.05	0.05	–	0.06	min
					ITERATION 3						
max	0.00	0.04	0.00	0.03	0.02	0.08	0.04	0.03	0.09	0.06	max
3	0.00	0.00	0.00	0.10	0.03	0.04	0.13	0.06	0.14	0.09	3
2	0.00	0.00	0.00	0.04	0.01	0.06	0.06	0.00	0.05	0.05	2
min	0.03	0.00	0.04	–	0.02	0.09	0.00	0.09	–	0.07	min

Note: To be read as "sanctions from ROW-contributor to COLUMN-contributor."[18]

that there is more punishment of players with higher contribution ranks by players with lower ranks in OPINION; however, this difference is clearly insignificant ($p = 0.2971$).

In the following, we will take a closer look at individual iterations separately. In iteration 1, the maximum-contributor punishes significantly more than other players without there being a treatment difference. There is a significant difference ($p = 0.0210$), however, with respect to the minimum-contributor. In BASIC, virtually no minimum-contributor ever carries out a punishment action in the first iteration, whereas in OPINION, this is roughly as likely as punishment by a player ranking second or third in terms of contributions.

In iteration 2, this difference between treatments diminishes since punishment activities of minimum-contributors in BASIC increase. Overall, there are no differences in punishment actions across treatments for any of the ranks ($p > 0.6$, all pair-wise comparisons), nor is there a difference in punishment between ranks within either treatment.

Interestingly, in the third iteration, the difference between BASIC and OPINION in terms of punishment activities by minimum-contributors reappears, although the difference between treatments is only weakly significant ($p = 0.0553$). So, although there is no general tendency for higher-contributing players to be punished more frequently by lower-contributing players in OPINION as pointed out earlier, there seems to be a specific

treatment difference concerning minimum-contributors. Given we do not have conclusive evidence on what may motivate this difference, we relegate its discussion to section 5.

The number of independent observations diminishes rapidly for most of the cells in iteration 3, so that little can be said due to a lack of data. However, there is an interesting point that comes to mind when eye-balling Table 2: the positive frequencies in the upper left-hand corner of the third-iteration tables could be a sign of sanction enforcement, in the sense of a player punishing another for not punishing a non-cooperative third (as, e.g., suggested by Henrich and Boyd 2001). However, the actions represented by these fractions are too few and can partially also be attributed to other potential explanations like "retarded" punishment actions. As a consequence, it is impossible to pin-point most of these actions as sanction enforcement.

4.3. **Contribution Norms**

4.3.1. Econometric Models.
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. As mentioned before, the number of instances of ongoing iterations beyond the third decreases rapidly. To rely on a sufficient number of observations, we 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 of coefficients 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, 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

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

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 five 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, 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, a group's average contribution and the punisher's own contribution, respectively. Models 4 and 5 test for an absolute norm.

Norm variables. For models 2–4 (5), we define two (one) distance measures each. For each of these models, we measure the absolute differences between the reference value under review and the contribution of the player to be punished, treating upward and downward deviations separately. The deviation terms are always defined by

$$\begin{aligned} n^- &:= |\min\{x_j^t - \tilde{x}^t, 0\}|, \text{ and} \\ n^+ &:= \max\{x_j^t - \tilde{x}^t, 0\}, \end{aligned} \tag{6}$$

where $\tilde{x}^t$ is the respective reference value, and n^- (n^+) denotes the corresponding downward (upward) deviation from this value. A summary of the models and their reference points is given in Table 3. Note that the variable n^- decreases in the punished player's contribution as long as this contribution is below the respective reference point. A significant positive

Table 3: Overview of the estimated norms

Variables: x_j^t is the punished player's contribution; $\overline{x}^t$ is the average contribution; x_i^t is the punisher's contribution; y is a constant integer number with $y \in [0, 20]$.

	Norm terms	Definition
Model 1	–	–
Model 2	$r_\star^-$	$\lvert\min\{x_j^t - \overline{x}^t, 0\}\rvert$
	$r_\star^+$	$\max\{x_j^t - \overline{x}^t, 0\}$
Model 3	$r_{\star\star}^-$	$\lvert\min\{x_j^t - x_i^t, 0\}\rvert$
	$r_{\star\star}^+$	$\max\{x_j^t - x_i^t, 0\}$
Model 4	a^-	$\lvert\min\{x_j^t - y, 0\}\rvert$
	a^+	$\max\{x_j^t - y, 0\}$
Model 5	a^-	$\lvert\min\{x_j^t - 20, 0\}\rvert$

effect of n^- would indicate that prosocial punishment is guided by the corresponding norm. If a norm determines antisocial punishment, we expect to find a significant positive effect of n^+.

Notice that for all norms we face another potential estimation result in terms of the norm coefficients: a positive coefficient for $r_\star^-$, $r_{\star\star}^-$, or a^-, implying that negative norm violations increase (the probability of) punishment, combined with a negative coefficient for $r_\star^+$, $r_{\star\star}^+$, or a^+, which would imply that a positive "norm violation" *decreases* the probability of punishment or the punishment level.[20] In this case, any deviation from contributing one's full endowment leads to an increase in the respective punishment determinant. In other words, subjects' elicited reference point would be nothing but their endowment, whereas the "norm term" merely identifies the location of a kink on the right-hand side of the probit equation. Given the scenario just described is exactly what we observe, we add the absolute-norm model with $y = 20$ as a fifth candidate to the models discussed. The difference between the log-likelihoods of this model 5 and the best-performing model will give us a first approximation of how much prediction power is lost by abstracting from the additional nonlinearity. This can, of course, only be treated as a rough estimate in light of the fact that the full-contribution model by its very nature exhibits a lower number of free parameters.

In all models that include one of the norms detailed above, we allow that specific norm to act differently in the two treatments. To incorporate this, we add an interaction effect between each norm part and a treatment dummy.

Two final remarks on our procedure seem warranted. 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 to increase our ability to distinguish between the absolute and the relative norms. In our presentation of the results, we select and report that absolute norm fitting the data best according to the log likelihood, based on a grid search over all possible contribution choices. 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.

Last but not least, notice that we retain the reference point of the punisher contribution 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

[20] Actually, there is yet another possibility, with $r_\star^-$, $r_{\star\star}^-$, and a^- coefficients being negative, and $r_\star^+$, $r_{\star\star}^+$, or a^+ coefficients being positive. This would mean that (the probability of) punishment increases in contributions, which, however, is rather counterintuitive and will not be discussed in the following section.

as a reference point (not reported here) is clearly outperformed by the reported model 3 on all iterations.

Controls. 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, Page, and Putterman 2006), whereas 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. In addition, 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 *RQ2*: 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) and others, including a second punishment stage in a public-good game may trigger severe retaliation. To analyze differences in retaliation across the two treatments, we include

the interaction effect $p^{t,1}_{j\to i} \times$ *opinion* ($p^{t,2}_{j\to i} \times$ *opinion*) in our regressions on announcements and on punishment points.

4.3.2. Estimation Results.

We organize our presentation of the results in the following way: first, we discuss the findings from our estimations on announcements and liken them to those on the assigned points. The discussion of potential treatment differences is deferred to a second step. Finally, we present the estimations with respect to voiced opinions, to account for the treatment differences in the level of point assignments.

In all regressions, an intercept is included, which, however, is not reported. We compare between the nested models (model 1 vs. models 2–5, respectively) on the basis of the Wald's χ^2-test. Asterisks indicate significance levels corresponding to this test. Other model comparisons are done on the basis of the test proposed by Vuong (1989). Unfortunately, for a majority of the comparisons, the test cannot be applied. In these instances, we have to rely on a comparison of the log-likelihoods which, as a consequence, only provides a tentative answer of which model to prefer.

Norm estimations. Results for the estimations of mean marginal effects on announcements are reported in Table 4, those for point assignments in Table 5. The most striking finding in terms of the focus of our paper is that in all iterations and (virtually) all models, our estimation results point to a full-contribution norm: on the one hand, the announcement probability as well as the amount of points assigned increase in downward deviations from the respective reference point as hypothesized, on the other, they *decrease* in upward deviations in all models in iterations 1 and 2 (often significantly, particularly in the best-performing models). In iteration 3, there is a single announcement model for which the corresponding coefficient is positive, even if insignificant (note that for point assignments, none of the reference-points contributes to explaining our data in this iteration). In other words, our estimation exercise *de facto* shows that the elicited reference point against which players' performance is measured is subjects' endowment in all iterations (but the third, for assignments). To summarize,

Result 1: *The probability of an announcement is determined by the distance between the punished players' endowment and their contribution.*[21] *Particularly, there is no reference value with the property that an increase in contributions above this value leads to an increase in the probability of being punished.*

In other words, empirically there is no apparent norm (apart from the full-contribution benchmark) that distinguishes "pro-social" and "anti-social"

[21] Research by Reuben and Riedl (2009) suggests that the determinant may be subjects' contribution capability rather than their endowment. Unfortunately, in our design the two cannot be discerned.

Table 4: Mean marginal effects for announcements

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
			Iteration 1		
x_i^t	0.0058***	0.003***	−0.002**	0.005***	0.005***
X_k^t	0.0014***	−0.0003	0.002***	0.002***	0.002***
round	−0.002***	0.001***	−0.001**	−0.001***	−0.001***
opinion	−0.012	0.012	0.003	0.007	0.038
$r_\star^-/r_{\star\star}^-/a^-$		0.017***	0.012***	0.01***	0.01***
$r_\star^+/r_{\star\star}^+/a^+$		−0.006	−0.005**	−0.009*	
$r_\star^-/r_{\star\star}^-/a^- \times op$		0.007	−0.004	−0.005	−0.005
$r_\star^+/r_{\star\star}^+/a^+ \times op$		−0.005	0.005	0.011	
Best absolute norm				15	20
Log likelihood	−1027.5	−801.5***	−798.3***,a	−809.4***	−813.5***
			Iteration 2		
x_i^t	0.001	0.0004	−0.0004	0.0008**	0.0009**
X_k^t	0.0003	0.00004	0.0004	0.0005*	0.0005*
round	−0.001***	−0.0006***	−0.0006***	−0.0005**	−0.0005***
opinion	−0.007	0.02	0.018	0.0012	0.033*
$p_{j\to i}^{t,1}$	0.0158***	0.018***	0.017***	0.018***	0.018***
$p_{j\to i}^{t,1} \times op$	0.0013	−0.008	0.005	−0.013	0.0186
$r_\star^-/r_{\star\star}^-/a^-$		0.004***	0.0027***	0.0026***	0.002***
$r_\star^+/r_{\star\star}^+/a^+$		−0.001	−0.0008	−0.0022*	
$r_\star^-/r_{\star\star}^-/a^- \times op$		−0.003	−0.0014	−0.0006	−0.002
$r_\star^+/r_{\star\star}^+/a^+ \times op$		−0.002	−0.0005	0.004	
Best absolute norm				10	20
Log likelihood	−383.8	−370.7***	−369.34***	−363.7***,b	−367.3***
			Iteration 3		
x_i^t	0.0001	0.0001	0.001	0.0001	0.0001
X_k^t	0.0007**	0.0006**	0.0006**	0.0006**	0.0006**
round	0.0003	0.0002	0.0003	0.0003*	0.0003
opinion	0.0072	0.008	0.0076	0.0081	0.022*
$p_{j\to i}^{t,2}$	0.012**	0.0104**	0.012**	0.011**	0.012**
$p_{j\to i}^{t,2} \times op$	0.0018	−0.0033	0.0021	0.0028	0.018
$r_\star^-/r_{\star\star}^-/a^-$		0.001*	0.0003	0.0012**	0.001**
$r_\star^+/r_{\star\star}^+/a^+$		−0.0009	−0.0005	0.0008	
$r_\star^-/r_{\star\star}^-/a^- \times op$		−0.002	−0.0001	−0.0006	−0.0019
$r_\star^+/r_{\star\star}^+/a^+ \times op$		0.0029	−0.0007	0.0067	
Best absolute norm				16	20
Log likelihood	−169.6	−166.3**	−168.6	−162.1**	−165.7**

Note: [a] ([b]) model fits significantly better than the second best model at $p < 0.1$ ($p < 0.05$), Vuong's test.

*** indicates significance at a $p < 0.01$ level; ** at a $p < 0.05$ level; * at a $p < 0.1$ level. Asterisks attached to log-likelihood values indicate the significance level of the Wald's χ^2-test comparing model 1 and the respective model.

Table 5: Mean marginal effects for points

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
			Iteration 1		
x_i^t	0.136^{***}	0.087^{***}	-0.0707^{*}	0.196^{***}	0.196^{***}
X_k^t	0.0445^{***}	-0.006	0.0803^{***}	0.1^{***}	0.1^{***}
round	-0.057^{**}	-0.029	-0.0287	-0.025	-0.028
opinion	-2.76^{***}	-0.465	-0.599	-5.59^{***}	1.43^{*}
sum_v^t	5.036^{***}	4.324^{***}	4.071^{***}	4.27^{***}	4.27^{***}
$r_{\star}^{-}/r_{\star\star}^{-}/a^{-}$		$0.597^{***,b}$	$0.428^{***,b}$	0.19	$0.40^{***,b}$
$r_{\star}^{+}/r_{\star\star}^{+}/a^{+}$		-0.359	-0.241^{**}	$-0.44^{***,b}$	
$r_{\star}^{-}/r_{\star\star}^{-}/a^{-} \times op$		$-0.599^{***,b}$	$-0.347^{***,b}$	-0.174	$-0.391^{***,b}$
$r_{\star}^{+}/r_{\star\star}^{+}/a^{+} \times op$		0.415	0.253^{**}	$0.431^{***,b}$	
Best absolute norm				3	20
Log likelihood	-1220.3	-1044.5^{***}	-1040.7^{***}	-1062.8^{***}	-1063.0^{***}
			Iteration 2		
x_i^t	0.084^{*}	0.054	-0.0232	0.118^{**}	0.123^{**}
X_k^t	0.0342	0.0051	0.0525	0.071	0.070
round	-0.082^{**}	-0.077^{**}	-0.072^{*}	-0.0686^{*}	-0.072^{*}
opinion	0.416	2.132	1.761	-0.535	4.05^{**}
sum_v^t	6.119^{***}	6.004^{***}	5.858^{***}	5.98^{***}	6.06^{***}
$p_{j\to i}^{t,1}$	1.777^{***}	2.286^{***}	2.422^{***}	2.379^{***}	2.379^{***}
$p_{j\to i}^{t,1} \times op$	-0.536	-1.05^{*}	-1.399^{**}	-1.11^{*}	-1.15^{*}
$r_{\star}^{-}/r_{\star\star}^{-}/a^{-}$		$0.503^{***,b}$	$0.357^{***,b}$	$0.392^{***,b}$	$0.35^{***,b}$
$r_{\star}^{+}/r_{\star\star}^{+}/a^{+}$		-0.158	-0.158	-0.287^{*}	
$r_{\star}^{-}/r_{\star\star}^{-}/a^{-} \times op$		$-0.482^{**,b}$	$-0.277^{*,b}$	-0.106^{b}	$-0.355^{**,b}$
$r_{\star}^{+}/r_{\star\star}^{+}/a^{+} \times op$		-0.059	0.176	0.472^{**}	
Best absolute norm				10	20
Log likelihood	-480.2	-468.1^{***}	-465.5^{***}	-462.5^{***}	-465.2^{***}
			Iteration 3		
x_i^t	-0.0206	-0.014	0.02	-0.03	-0.03
X_k^t	0.077^{***}	0.0784^{***}	0.0654^{***}	0.069^{***}	0.069^{***}
round	0.1^{***}	0.105^{***}	0.105^{***}	0.114^{***}	0.110^{***}
opinion	-2.14^{**}	-2.057^{**}	-2.191^{**}	-2.334^{*}	-1.328
sum_v^t	4.197^{***}	4.783^{***}	4.571^{***}	4.409^{***}	4.559^{***}
$p_{j\to i}^{t,2}$	-3.705^{***}	-3.587^{***}	-3.569^{***}	-3.501^{***}	-3.53^{***}
$p_{j\to i}^{t,2} \times op$	4.319^{***}	4.363^{***}	4.299^{***}	4.222^{***}	4.245^{***}
$r_{\star}^{-}/r_{\star\star}^{-}/a^{-}$		-0.0183^{a}	-0.014^{a}	0.0085	0.0026^{a}
$r_{\star}^{+}/r_{\star\star}^{+}/a^{+}$		-0.0285	0.0081	0.0385	
$r_{\star}^{-}/r_{\star\star}^{-}/a^{-} \times op$		$-0.346^{*,a}$	$-0.178^{*,a}$	-0.1065	$-0.176^{*,a}$
$r_{\star}^{+}/r_{\star\star}^{+}/a^{+} \times op$		0.095	0.065	0.395	
Best absolute norm				17	20
Log likelihood	-343.6	-339.1	-339.9	-338.5	-339.1

Note: ^{a} (^{b}) the sum of the norm and the interaction between the norm and OPINION equals zero at $p < 0.1$ ($p < 0.05$), F-test.

*** indicates significance at a $p < 0.01$ level; ** at a $p < 0.05$ level; * at a $p < 0.1$ level. Asterisks attached to log-likelihood values indicate the significance level of the Wald's χ^2-test comparing model 1 and the respective model.

or "perverse" punishment. If "perverse" punishment was norm-related behavior, there is no sign of it in our data.

The second main finding is that the application of the full-contribution standard differs between iterations. This can be seen from the fact that in iteration 1, model 3 performs best in all decision contexts (with a weakly significant difference to the next-best model for announcements), but that it is outperformed by absolute-norm models in subsequent iterations. For both announcements and point assignments (and opinions, but more on that later), behavior in the first iteration is modulated strongly by the punisher's contribution. Although the reference standard for who should be punished is (the punished) players' endowment as we have seen, the trigger for a punishment action often seems to be the potential punisher's contribution relative to that of the player to be punished. An intuitive explanation that has been proposed in the literature is that high-contributors do not want to be the "sucker" (e.g., Fehr and Gächter 2000, Burlando and Guala 2005). The larger the difference between the two players' contributions, the stronger the emotional response (e.g., Fehr and Gächter 2002, Xiao and Houser 2005), and therefore, the more likely punishment is triggered. However, once the first iteration is over, the importance of the punisher's relative contribution wears away. There may still be some second-order nonlinearity with respect to the punished player's contribution level—as indicated by the fact that the best fit for models of type 4 is achieved for a y of 10 (iteration 2) and 16 (iteration 3, announcements; for assignments, $y = 17$)—but generally not much is to be gained by splitting the full-contribution norm of model 5.

Result 2: *In the first iteration, the announcement of punishment is accentuated by the punisher's contribution relative to that of the punished player. This difference in contributions also influences strongly the level of punishment. In later iterations, this is no longer the case.*

Let us shortly review the effects of our control variables that, by and large, have the effects one might expect. The punisher's absolute contribution level has a positive effect on both announcements and point assignments, as do the contributions of the players who are neither the punisher nor the target of the punishment action[22]; the likelihood of an announcement decreases in the course of the experiment, as does the punishment

[22] 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 punished player'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 punished player's contributions *for a given distance*. Thus, increasing both the punisher's and the punished player's contributions by the same amount decreases both announcements and point assignments (this finding points at the absolute nature of the contribution norm). 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 earlier.

level in iterations 2 and 3; finally, the number of punishment points received in the preceding iterations is a strong indicator for both the likelihood and the level of punishment actions in iterations 2 and 3. Interestingly, in iteration 3, punishment points received have a *negative* impact on punishment assignments ($p^{t,2}_{j\to i}$ in Table 5). A tentative explanation for this may be that, although subjects do not want to "give in," they do start to economize on resources in this iteration, potentially in order not to nullify their round earnings completely.

Treatment effects. The first thing to note is that for announcements, none of the interaction variables across all models and iterations turns out to be significant. Furthermore, only model 5 points to a weakly significant treatment dummy in iterations two and three. There is no such effect in iteration 1, and none of the second-order nonlinear models exhibits the effect in any iteration. In view of the above, we conclude that announcements in the two treatments seem to be governed by the same rules.

This finding contrasts sharply with what we observe in terms of assigned points. In many of the models for iterations 1 and 2 (including those performing best in terms of the log-likelihood), the norm coefficients in both treatments are significantly different as evidenced by the significant interaction effects. More surprisingly, *F*-tests provide statistical evidence that in all of these models, at least one of the interaction effects exactly cancels out the corresponding norm effect. In all cases, the coefficient of the norm-interaction effect bears the opposite sign of the norm-effect coefficient. Even in those cases where an *F*-test does not signalize statistical significance, the opposed effect sizes are of a similar magnitude. In iteration 3, the statistical significance is much weaker but the central tendency stays the same.

What the above result seems to suggest is that in OPINION, the punished players' contribution does not have a (direct) effect on the level of punishment points assigned. Instead, the sum of votes takes over the role of main determinant, as can be seen from the significance of the sum^t_v coefficient. It could, of course, be argued that the number of points assigned and the number of favorable opinions could simply be perfectly correlated, as the severity of a player's misbehavior could lead independently to both greater approval and stronger punishment, without one affecting the other. Because we cannot use an instrumental-variable approach in our design, we cannot claim that this is a causal interference. Yet, the severity of misbehavior is exactly what the norm terms should capture, so that it seems safe to speak of a reinforcing effect of social approval on the punishment level.

To learn more about the characteristics of the way the voiced opinions are formed, we apply the same type of analysis we used for announcement probabilities and assignment levels to the probability of voicing a favorable opinion. Before we do so, we would like to point out some interesting figures on retaliation and opinions thereupon. About 25% of all announcement in iteration 2 can be clearly classified as retaliation; of all opinions voiced on these announcements, 23% are positive. In iteration 3, 25% of

all announcement can be classified as retaliation. Compared to iteration 1, the fraction of positive evaluations more than doubles, to 50%. At the same time, re-retaliation (only 3% of all announcements in iteration 3) is never approved of. Notice that of all second-iteration counter-punishers, only about 8% received punishment by a different player in iteration 3. In other words, counter-punishment—when it happens—does not seem to be unacceptable *per se*.[23]

The above facts suggest bystanders have a very differentiated picture on retaliative punishment. To shed light on this issue, we separate $p^{t,1}_{j\to i}$ ($p^{t,2}_{j\to i}$, respectively) in iterations two and three in two dimensions: into mild and harsh sanctions, assigned to high and low contributors, respectively.[24] Specifically, let $m^{t,1}_{j\to i}$ ($m^{t,2}_{j\to i}$) measure the number of punishment points player i receives from j in the first (second) iteration, if at most two points were assigned (that is, at most the median number of assigned punishment points). Similarly, let $h^{t,1}_{j\to i}$ ($h^{t,2}_{j\to i}$) measure the number of punishment points player i receives from j in the first (second) iteration, if at least three points were assigned. Therefore, the inclusion of $h^{t,1}_{j\to i}$ and $h^{t,2}_{j\to i}$ ($m^{t,1}_{j\to i}$ and $m^{t,2}_{j\to i}$) allows to test whether the probability of a favorable opinion changes depending on whether the announcing player received harsh (mild) punishment in the preceding iteration. On a second dimension, we distinguish contribution types: let $H = 0$ if i contributed no more than 10 to the public good (that is, half of the contribution norm in the previous iteration), and $H = 1$ otherwise. Then, significance of the interaction effects $H \times h^{t,1}_{j\to i}$ and $H \times m^{t,1}_{j\to i}$ ($H \times h^{t,2}_{j\to i}$ and $H \times m^{t,2}_{j\to i}$, respectively) would indicate that bystanders judge retaliation by high-contributing players i differently to retaliation by low contributors.[25] The results of our analysis are presented in Table 6.

In our estimation of the norms governing bystanders' opinions about punishment actions, we observe a pattern that is rather similar to those obtained for announcements and point assignments: in iteration 1, model 3 seems to perform better than its competitor models, whereas in iterations 2 and 3, the advantage is on model 4's side. The former points to a reference point equal to subjects' endowment, as would model 4 in iteration 2, if we were to judge by the norm-related coefficients even though they are not significantly different from zero. The fact that model 5, having one less free parameter does not perform substantially worse while exhibiting a significant norm-coefficient seems to give some backing to this claim. The actual surprise happens in iteration 3, where we observe the only instance of a

[23] If we pool data from BASIC and OPINION, the corresponding numbers do not differ substantially: about 32% (26%) of all announcement in iteration 2 (3) are retaliation, 8% of all announcements in iteration 3 are re-retaliation, whereas about 7% of second iteration counter-punisher receive third-party punishment in the third iteration.

[24] We are grateful to an anonymous referee for suggesting this more nuanced analysis.

[25] Notice that taking into account the interaction between punishment severeness and the punished player j's contribution for approval does not yield additional evidence.

Table 6: Mean marginal effects for opinions

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
			Iteration 1		
x_i^t	0.0004	-0.0329^{***}	-0.0776^{***}	-0.0102	-0.0100
x_k^t	0.0115^{*}	0.0146^{**}	0.0368^{***}	0.0380^{***}	0.0373^{***}
round	-0.0014	-0.0011	-0.0015	-0.0012	-0.0011
$r_\star^-/r_{\star\star}^-/a^-$		0.0725^{***}	0.0545^{***}	0.0642^{***}	0.0607^{***}
$r_\star^+/r_{\star\star}^+/a^+$		-0.0999^{***}	-0.0848^{***}	0.0337	
Best absolute norm				19	20
Log likelihood	-196.0	-144.3^{***}	-134.4^{***}	-135.4^{***}	-135.6^{***}
			Iteration 2		
x_i^t	-0.0053	-0.0073^{*}	-0.0120^{**}	-0.0039	-0.0048
x_k^t	0.0008	0.0017	0.0044	0.0041	0.0044
round	-0.0005	-0.0007	-0.0006	-0.0004	0.0005
$m_{j\to i}^{t,1}$	$-0.5664^{**,b}$	$-0.4495^{***,b}$	$-0.4236^{**,b}$	$-0.4057^{**,b}$	$-0.4235^{**,b}$
$H \times m_{j\to i}^{t,1}$	$0.6070^{***,b}$	$0.4884^{***,b}$	$0.4601^{**,b}$	$0.4390^{**,b}$	$0.4600^{**,b}$
$h_{j\to i}^{t,1}$	-0.0002	0.0071	0.0075	0.0080	0.0080
$H \times h_{j\to i}^{t,1}$	0.0351^{**}	0.0268	0.0219	0.0171	0.0212
$r_\star^-/r_{\star\star}^-/a^-$		0.0114^{*}	0.0074^{**}	0.0028	0.0073^{**}
$r_\star^+/r_{\star\star}^+/a^+$		-0.0038	-0.0069	-0.0096^{**}	
Best absolute norm				9	20
Log likelihood	-42.2	-36.5^{**}	-35.3^{**}	-35.0^{***}	-35.3^{***}
			Iteration 3		
x_i^t	-0.0084	0.0002	-0.0583^{**}	0.0101	0.0059
x_k^t	-0.0018	-0.0012	0.0032	-0.0021	0.0043
round	0.0088	0.0042	0.0020	0.0050	0.0020
$m_{j\to i}^{t,2}$	0.0624	0.1423	-0.1776	$-0.4021^{***,a}$	-0.1520
$H \times m_{j\to i}^{t,2}$	-0.1050	-0.1486	0.1817	$0.4352^{***,a}$	0.1539
$h_{j\to i}^{t,2}$ [c]	0.0515	-0.0474	-0.1817	-0.0284	-0.0198
$r_\star^-/r_{\star\star}^-/a^-$		0.1681^{***}	0.0556^{***}	0.1975^{***}	0.0605^{***}
$r_\star^+/r_{\star\star}^+/a^+$		-0.0038	-0.0674^{**}	0.0424^{*}	
Best absolute norm				15	20
Log likelihood	-41.9	-30.9^{***}	-33.7^{***}	-30.2^{***}	-33.8^{***}

Note: [a] ([b]) the sum of received punishment points and the interaction between the sum and dummy viable for above 10 contributors equals zero at $p < 0.1$ ($p < 0.05$), F-test; [c] $H \times h_{j\to i}^{t,2}$ is not applicable because we have only two observations.
*** indicates significance at a $p < 0.01$ level; ** at a $p < 0.05$ level; * at a $p < 0.1$ level. Asterisks attached to log-likelihood values indicate the significance level of the Wald's χ^2-test comparing model 1 and the respective model.

reference point that is clearly different from the full contribution of 20. Surprisingly, punishment acts directed at players contributing more than three quarters of their endowment are applauded significantly more, the higher those players' contribution was. The fact that the corresponding model is

the only model to (highly) significantly outperform model 1 would suggest that something has drastically changed in the way players evaluate other players' actions in iteration 3. However, looking at the data more closely, we note that most (12 out of 17) punishment actions directed at players contributing more than 15 (the estimated absolute norm) stem from the same matching group. The small number of observations of such behavior outside the mentioned matching group casts some doubt on the robustness of the reported finding.[26] If this effect also arises in future studies comprising more observations, it poses a serious challenge for the scholarly community, as there is no obvious reason for why the determinants of bystanders' opinions should change after two iterations.

Let us look at the iterations in a little more detail. In iteration 1, favorable opinions are more frequent the higher the contributions of the player voicing his opinion ($x_k^t > 0$ in all models). With respect to the punisher, the evidence is inconclusive. The worse-performing model 2 indicates a negative influence of punisher contributions, whereas models 1, 4, and 5 do not find evidence for opinions being influenced by the punisher's contribution. Model 3, finally, suggests a positive influence of the punisher contribution (cf. footnote 22 for the argument), on top of the general full-contribution reference point indicated by the significantly negative coefficient of r_{**}^{+} as well as by the good performance of model 5. Put differently, similar to our findings for punisher decisions, the best-performing model 3 suggests a modulation of a full-contribution norm by the punisher's contribution.

The results of iteration 2 are similar to those of iteration 1, except for the interesting fact that punishers' and especially bystanders' contributions lose (much of) their influence. This finding is remarkable: players' opinion about a punishment action seems to be independent of their level of cooperativeness. Instead, bystanders pay attention to the punishment history: the frequency of favorable opinions is significantly lower for mildly punished players who contributed no more than half of the endowment than for high-contributing players. In other words, when subjects judge retaliative actions, they seem to differentiate between high and low contributors to the public good: favorable opinions are significantly less frequent for retaliation by low-contributing players (notice that the F-test yields $p < 0.05$ for the sum of $m_{j\to i}^{t,1}$ and $H \times m_{j\to i}^{t,1}$ equaling zero, cf. Table 6). There is only weak evidence that harshly sanctioned players are applauded when defending themselves: while the coefficients for $h_{j\to i}^{t,1}$ and $H \times h_{j\to i}^{t,1}$ are consistent across models 2–5, they are not significant in any of these models.[27] On the other hand, in models

[26] Of course, this robustness issue in some sense extends beyond the effects on opinions; however, for announcements and point assignments, we combine the data from both treatments, which substantially increases the number of observations used in the estimation process.

[27] This may, of course, be due to the fact that most observations for first-round punishment fall into the "mild category."

2–5, the norm continues to play a significant role in the expected direction (in model 4, this is reflected in the significantly negative coefficient for a^+). The best-performing model indicates the absolute nature of the norm. The small performance difference between models 4 and 5 indicates that the kink at a contribution of 9 (model 4) does not add much explanatory power. We interpret this as further evidence for the full-contribution norm.

In iteration 3, the general picture has changed slightly: only the best-performing model 4 indicates a differentiation between retaliating players according to their contributions. According to this model, low-contributing retaliators receive less favorable opinions than other punishers. At the same time, the model favors an absolute norm that differs from what we have seen in all other iterations and for punishment decisions in the same iteration, as discussed earlier.

Having analyzed punishment decisions and bystanders' opinions in detail, we step back to take a look at the broad picture: overall, we have the astonishing result that there are few differences between punishment behavior in BASIC and OPINION, even though the behavior in OPINION is determined by a different data-generating process than in BASIC. However, because voiced opinions are based on the same criteria as punishment announcements in either treatment and as punishment severity in BASIC, we do not observe behavioral differences between the two treatments. Abstracting from the surprising third-iteration effect for opinions, we are ready to answer research question *RQ 3* affirmatively:

Result 3: *The full-contribution norm guiding punishment actions and social approval is the same, even though the mechanism by which the norm determines punishment severity differs between treatments.*

There are, furthermore, differences with respect to its application. The existence of an opinion poll seems to dilute the effect that the presence of multiple punishment stages seems to have, namely that lower-contributing players do not punish high contributors in the first iteration. The dependence of punishment assignments on the social opinion seems to compensate for this at least partially, accentuating the importance of the full-contribution norm in what looks like a re-focusing way. The effect of this mechanism is punishment patterns in the two treatments that are barely distinguishable.

5. Discussion

In a recent study, Carpenter and Matthews (2009) find 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 find for the decision on whether to assign punishment points and that on how many to assign differ 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 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 four 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; (iii) we contribute to understanding behavior under the realistic assumption that punished players may retaliate, that studies like Denant-Boemont, Masclet, and Noussair (2007) and Nikiforakis (2008) have shown to lead to substantially different behavior from what is usually observed in public-good experiments with peer punishment (as in Fehr and Gächter 2000); and (iv) we take our examination one step further than the above counter-punishment studies by removing the arbitrariness of a prespecified number of punishment stages. 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 preceding studies.

Our findings are noteworthy in a number of ways. First of all, contributions and earnings are fairly stable over time, with round-wise earnings being above the Nash-equilibrium level of 20 tokens. In other words, the introduction of counter-punishment opportunities does not lead to a breakdown of cooperation. This is a notable difference with respect to the data reported in Denant-Boemont, Masclet, and Noussair (2007) and Nikiforakis (2008) but closely corresponds to the findings of Nikiforakis and Engelmann (2011), which is closest in design to our study. There may be a number of reasons for this difference. In Nikiforakis (2008) and all but the "6FSI" treatment of Denant-Boemont, Masclet, and Noussair, there is exactly one retaliation stage. This means that punishers who sanction anti-social behavior face the threat of counter-punishment without being able to respond to it. This may discourage first-order punishment. In designs with more than two punishment stages, retaliators against 'warranted' punishment face the threat of being retaliated against, or even of facing third-order punishment by others. This may be expected to substantially decrease retaliation, which may in turn

(re-)encourage first-order punishers. If this explanation is valid, we are only left with one difference unaccounted for, namely that between the "6FSI" treatment of Denant-Boemont, Masclet, and Noussair on the one side, and our results and those of Nikiforakis and Engelmann (2011), on the other. One possible explanation would be the following: in the "6FSI" treatment, there is a prespecified number of six consecutive stages in which players can assign punishment points. This means that punishment might be deferred to the sixth stage to prevent retaliation from taking place. However, the data reported in Denant-Boemont, Masclet, and Noussair (2007, p. 156) shows this is not the case in general: instead, there is extraordinarily heavy punishment already in the first punishment stage. At the same time, Denant-Boemont, Masclet, and Noussair (2007, p. 164) report "a large increase [in stage-wise punishment assignments] in stage 6." In other words, their setup has the mentioned punishment-deferring effect, but this effect seems to apply only partially. What this may suggest is that early punishers try to discourage (deferred) retaliation by sharply reducing the sanctioned player's period-earnings straight away. If this is part of what happens in their data, then the "6SFI" treatment has important aspects in common with single-retaliation-stage setups that are not present in studies with an endogenous number of stages like Nikiforakis and Engelmann (2011) or our own.

With respect to our research interest in cooperation norms, we find support for a finding already made by Carpenter and Matthews: the average-related contribution norm, being the focus of a non-negligible number of studies is outperformed as a predictor of behavior by other models in every iteration and for each decision. Thus, our data supports the development in recent studies to depart from the assumption of the average contribution being the norm (e.g., Herrmann, Thöni, and Gächter 2008, Egas and Riedl 2008).

Furthermore, like Carpenter and Matthews, we find strong support for the influence of an absolute cooperation norm. This norm is subjects' full endowment which—in contrast to the findings of Carpenter and Matthews (2009)—is consistent over decisions, iterations, and roles (punisher or bystander). This lends support to the argument brought forward by Fehr and Schmidt (1999) to select the full-contribution equilibrium as being focal.[28] What these models of prosocial behavior do not account for is the fact that players are also willing to punish those who contribute the same or even more than they do—yet in a *prosocial* fashion, that is, because these others still contribute less than the full-contribution norm. This is a challenge our results pose for future theories of social behavior.

[28] As an anonymous referee correctly points out, it should be remembered that these claims are limited to the case of linear public goods. It is conceivable that in case of an interior social optimum, the reference point would be given by the symmetric contribution leading to that optimum rather than by subjects' endowment.

Even under the full-contribution norm, subjects are prompted to increase both the punishment probability and its severity in the first iteration if the player to be punished has deviated more from the full-contribution norm than the punisher him or herself. This suggests a high level of personal investment in the public-good dilemma, triggering strong emotional responses as already suggested by studies such as Fehr and Gächter (2002) or Xiao and Houser (2005). The effect vanishes in higher iterations, suggesting that others' contribution levels are judged in a more objective manner. At the same time, retaliation steps in as an important motive for punishment. This happens only in iteration 2, as we successfully eliminate the need for first-iteration "pre-emptive counter-punishment" by introducing multiple-stage punishment.

With respect to the bystanders' elicited norm, we find that in the first iteration, bystanders follow the same criteria as punishers in their announcement decisions, corroborating the claim that we are, indeed, facing a social norm. In the second iteration, we find further evidence for the full-contribution norm, even though the importance given to the players' contributions vanishes more rapidly than in the case of punishers' announcement decisions. Most interestingly, the bystanders' own contribution seems to play a role only in the first iteration. This seems to suggest that in iteration 1, high-contributors "vote with" high-contributing punishers (in the sense of seconding their announcement) and low-contributors "vote with" low-contributing players who are subject to punishment (in the sense of not favoring the latter punishment), whereas in later iterations, opinions seem to be of a more impartial nature. This seems to correspond well with the effect hypothesized for punishment decisions, namely that there is a high level of personal investment in the public-good dilemma only in iteration 1. In later iterations, the focus of both punishers and bystanders seems to move away from the dilemma and center on whether a punishment action is justified and appropriate. With respect to the question of whether retaliative actions are justified, there is a noteworthy difference between high- and low-contributors: unfavorable opinions are significantly more frequent for retaliation of low-contributing players than for high contributors. Unexpectedly, our results indicate a shift from a full-contribution norm to one of three quarters of the endowment in iteration 3. This shift is surprising and unaccounted for. At the same time, it has to be noted that the corresponding regression is on data from the third iteration of the OPINION treatment only, so that the effect may not prove to be robust in future studies. If it did, this effect would pose a serious challenge to any theory trying to account for the observed data.

Finally, our treatment variation does not seem to change behavior in a substantial way. However, there are a couple of remarkable treatment differences. First of all, in OPINION the punished players' contribution does not seem to have any effect on the severity of punishment. Rather, it is the number of favorable opinions that is the main determinant of the punishment

level.[29] And second, in contrast to the findings from our BASIC treatment, players in OPINION do seem to engage in a form of 'pre-emptive counter-punishment.' Although this finding seems counterintuitive at first sight, there may be a simple explanation for it. We have seen that a punisher's decision to punish a certain player depends on that player's contribution, and that bystanders tend to follow similar criteria. Therefore, a minimum-contributor having contributed very little faces a high probability of being punished *and* a likely endorsement of this punishment action by 'society'. However, our analysis has also shown that higher endorsement leads to substantially higher punishment levels. In other words, our minimum-contributor faces the threat of being punished much more severely in the OPINION treatment due to social endorsement, which, in a sense, bears resemblance to mob law. In this situation, a minimum-contributor may try to issue a warning by announcing a punishment action against the player she thinks to be the most likely punisher, which would result in the observed pattern. This explanation is, of course, only tentative speculation.

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.

Appendix A: Instructions[30]

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

[29] Again, we would like to stress that our design does not allow the use of an instrumental-variable approach, so that our claim is not a causal inference. The number of points assigned and the number of favorable opinions could simply be strongly correlated. Nonetheless, it is a very interesting observation that calls for future research.

[30] 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.

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.

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 × (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 ex-

actly 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. In addition, 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.

(i) 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?

(ii) 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?

(iii) 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?

(iv) 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?

(v) 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?

(vi) What are the total costs if you announce to distribute points to one of the group members and distribute 0 points?

(vii) What is the reduction of first stage income if you receive in total

- 0 points
- 4 points
- 15 points

from the other group members?

(viii) 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?

References

ANDERSON, C. M., and L. 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.

BURLANDO, R. M., and F. GUALA (2005) Heterogeneous agents in public goods experiments, *Experimental Economics* **8**, 35–54.

CARPENTER, J. P., and P. H. MATTHEWS (2009) What norms trigger punishment? *Experimental Economics* **12**, 272–288.

CINYABUGUMA, M., T. PAGE, and L. PUTTERMAN (2006) On perverse and second-order punishment in public goods experiments with decentralized sanctioning, *Experimental Economics* **9**, 265–279.

DAWES, C. T., J. H. FOWLER, T. JOHNSON, R. McELREATH, and O. SMIRNOV (2007) Egalitarian motives in humans, *Nature* **446**, 794–796.

DENANT-BOEMONT, L., D. MASCLET, and C. NOUSSAIR (2007) Punishment, counterpunishment and sanction enforcement in a social dilemma experiment, *Economic Theory* **33**, 145–167.

EGAS, M., and A. RIEDL (2008) The economics of altruistic punishment and the maintenance of cooperation, *Proceedings of the Royal Society B: Biological Sciences* **275**, 871–878.

FALK, A., and U. FISCHBACHER (2006) A theory of reciprocity, *Games and Economic Behavior* **54**, 293–315.

FEHR, E., and S. GÄCHTER (2000) Cooperation and punishment in public goods experiments, *American Economic Review* **90**, 980–994.

FEHR, E., and S. GÄCHTER (2002) Altruistic punishment in humans, *Nature* **415**, 137–150.

FEHR, E., and K. SCHMIDT (1999) A theory of fairness, competition, and cooperation, *Quarterly Journal of Economics* **114**, 817–868.

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

FISCHER, S., and A. NICKLISCH (2007) Ex interim voting: An experimental study of referendums for public good provision, *Journal of Institutional and Theoretical Economics* **163**, 56–74.

GÄCHTER, S., and B. 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.

GREINER, B. (2004) An online recruitment system for economic experiments. In: *Forschung und wissenschaftliches Rechnen 2003: GWDG Bericht* **63**, K. Kremer and V. Macho, eds., pp. 79–93. Göttingen: Gesellschaft für Wissenschaftliche Datenverarbeitung.

HENRICH, J., and R. 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.

HERRMANN, B., C. THÖNI, and S. GÄCHTER (2008) Antisocial punishment across societies, *Science* **319**, 1362–1367.

ISAAC, R. M., K. F. McCUE, and C. R. PLOTT (1985) Public good provision in an experimental environment, *Journal of Public Economics* **26**, 51–74.

JOHNSON TIM, C. T. D., J. H. FOWLER, R. McELREATH, and O. SMIRNOV (2009) The role of egalitarian motives in altruistic punishment, *Economics Letter* **102**, 192–194.

KROLL, S., T. L. CHERRY, and J. F. SHOGREN (2007) Voting, punishment, and public goods, *Economic Inquiry* **45**, 557–570.

MASCLET, D., C. N. NOUSSAIR, S. TUCKER, and M.-C. VILLEVAL (2003) Monetary and nonmonetary punishment in the voluntary contributions mechanisms, *American Economic Review* **93**, 366–380.

MASCLET, D., C. N. NOUSSAIR, and M.-C. VILLEVAL (2009) Threat and punishment in public good experiments. Working Paper.

MARGREITER, M., M. SUTTER, and D. DITTRICH (2005) Individual and collective choice and voting in common pool resource problems with heterogeneous actors, *Environmental and Resource Economics* **32**, 241–271.

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

NIKIFORAKIS, N., and D. ENGELMANN (2011) Altruistic punishment and the threat of feuds, *Journal of Economic Behavior & Organization* **78**, 319–332.

NOUSSAIR, C., and S. TUCKER (2007) Public observability of decisions and voluntary contributions in a multiperiod context, *Public Finance Review* **35**, 176–198.

ONES, U., and L. PUTTERMAN (2007) The ecology of collective action: A public goods and sanctions experiment with controlled group formation, *Journal of Economic Behavior and Organization* **62**, 495–521.

OSTROM, E. (2000) Collective actions and the evolution of social norms, *The Journal of Economic Perspectives* **14**, 137–158.

OSTROM, E., J. M. WALKER, and R. GARDNER (1992) Covenants with and without a sword: Self-governance is possible, *The American Political Science Review* **86**, 404–417.

REGE, M., and K. TELLE (2004) The impact of social approval and framing on cooperation in public good situations, *Journal of Public Economics* **88**, 1625–1644.

REUBEN, E., and A. RIEDL (2009) Enforcement of contribution norms in public good games with heterogeneous populations. Working Paper.

SEFTON, M., R. SHUPP, and J. WALKER (2007) The effect of rewards and sanctions in provision of public goods, *Economic Inquiry* **45**, 671–690.

SETHI, R. (1996) Evolutionary stability and social norms, *Journal of Economic Behavior and Organization* **29**, 113–140.

SOBER, E., and D. S. WILSON (1998) *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

SUGDEN, R. (1986) *The Economics of Rights, Co-operation, and Welfare*. Oxford: Blackwell Publishing Limited.

SUTTER, M., S. HAIGNER, and M. G. KOCHER (2010) Choosing the stick or the carrot? Endogenous institutional choice in social dilemma situations, *Review of Economic Studies* **77**, 1540–1566.

VUONG, Q. H. (1989) Likelihood-ration tests for model selection and non-nested hypotheses, *Econometrica* **57**, 307–333.

WALKER, J., R. GARDNER, A. HERR, and E. OSTROM (2000) Collective choice in the commons: Experimental results on proposed allocation rules and votes, *Economic Journal* **110**, 212–234.

XIAO, E., and D. HOUSER (2005) Emotion expression in human punishment behavior, *Proceedings of the National Academy of Sciences* **102**, 7398–7401.

YAMAGISHI, T. (1986) The provision of a sanctioning system as a public good, *Journal of Personality and Social Psychology Review* **51**, 110–116.

ZELMER, J. (2003) Linear public goods experiments: A meta-analysis, *Experimental Economics* **6**, 299–310.