



# Monopolizing Sanctioning Power under Noise Eliminates Perverse Punishment But Does Not Increase Cooperation

Sven Fischer<sup>1\*</sup>, Kristoffel Grechenig<sup>2</sup> and Nicolas Meier<sup>3</sup>

<sup>1</sup> Newcastle University, Newcastle University Business School, Newcastle upon Tyne, UK, <sup>2</sup> Max Planck Institute for Research on Collective Goods, Bonn, Germany, <sup>3</sup> School of Business and Economics, RWTH Aachen University, Aachen, Germany

We run several experiments which allow us to compare *cooperation* under *perfect* and *imperfect information* in a *centralized* and *decentralized punishment* regime. Under perfect and extremely noisy information, aggregate behavior does not differ between institutions. Under intermediate noise, punishment escalates in the decentralized peer-to-peer punishment regime which badly affects efficiency while sustaining cooperation for longer. Only decentralized punishment is often directed at cooperators (perverse punishment). We report several, sometimes subtle, differences in punishment behavior, and how contributions react.

**Keywords:** cooperation, public good, centralized punishment, imperfect information, anti-social punishment, perverse punishment

## OPEN ACCESS

### Edited by:

Benedikt Herrmann,  
University of Nottingham, UK

### Reviewed by:

Mark Bernard,  
Goethe University Frankfurt, Germany  
Daniele Nosenzo,  
University of Nottingham, UK

### \*Correspondence:

Sven Fischer  
sven.fischer@newcastle.ac.uk

**Received:** 01 December 2014

**Accepted:** 09 September 2016

**Published:** 29 September 2016

### Citation:

Fischer S, Grechenig K and Meier N  
(2016) Monopolizing Sanctioning  
Power under Noise Eliminates  
Perverse Punishment But Does Not  
Increase Cooperation.  
*Front. Behav. Neurosci.* 10:180.  
doi: 10.3389/fnbeh.2016.00180

## 1. INTRODUCTION

Modern societies have centralized sanctioning power as a means to enforce norms (Weber, 1919). This monopoly has often been justified on the premise that private, decentralized enforcement has (higher) negative externalities (Clotfelter, 1978; Polinsky, 1980). But, is centralization necessarily better? Experiments on voluntary cooperation repeatedly demonstrate that, compared to an environment *without punishment*, decentralized, informal, peer-to-peer punishment increases cooperation under perfect information (Yamagishi, 1986; Ostrom et al., 1992; Fehr and Gächter, 2000, 2002), and welfare in the long run (Gächter et al., 2008). Various studies, however, challenge the idea that peer-to-peer punishment generally enhances cooperation (for an overview, see Nikiforakis, 2014): for example on the basis of punishment which is targeted at cooperators, referred to as anti-social (Herrmann et al., 2008) or perverse punishment (Cinyabuguma et al., 2006), or on the basis of counter-punishment (targeted either at the group or the punisher directly as in Nikiforakis, 2008; Nikiforakis et al., 2012)<sup>1</sup>. If punishment is centralized, however, there is no opportunity for counter punishment and such sources of inefficiency are less likely. Furthermore, with only one punisher, there are no coordination problems and no problems resulting from possibly conflicting contribution norms. Some experiments test the effectiveness of *formal, centralized* enforcement mechanisms compared to informal, decentralized regimes, while capturing important aspects of institutions. This literature characterizes centralization as a mechanism that allows to commit to a sanctioning scheme, such that punishment is automatically carried out and/or determined according to some exogenous voting rule (Kosfeld and Riedl, 2004; Tyran and Feld, 2006; Guillen et al., 2007; Kube and Traxler, 2011; Putterman et al., 2011; Andreoni and Gee, 2012; Ambrus and Greiner, 2015). Centralization is viewed as a commitment mechanism

and these papers do not explore the effect of centralization *per se*, i.e., of merely concentrating all power into one hand.

In an environment with a centralized punisher rather than an automated mechanism, an important question is how the authorization to punish is granted. Baldassarri and Grossman (2011) compare two centralized punishment regimes. In one, authority is granted by chance, in the other by electing the punisher from within the group after two trial rounds. While punishment behavior is very similar, in the elected authority treatment participants cooperate more. This suggests that in order to test whether centralization *per se* matters, one needs to abstract from any mechanism that grants legitimacy, and restrict the analysis to a fair random selection. There are a few studies which allow to compare a centralized with a decentralized punishment regime in this way, and the resulting evidence is mixed. In Carpenter et al. (2012), where the role of the punisher is randomly allocated but fixed and punishment is cheap, contributions and overall efficiency are larger in the decentralized regime. Similarly, in Nosenzo and Sefton (2014) contributions are substantially larger in a mutual punishment regime than in a centralized one. In O’Gorman et al. (2009), on the other hand, where authority is also granted at random but changes every round, and where punishment is expensive, there are no significant differences in cooperation. However, due to significantly more punishment, the decentralized regime is less efficient.

One conclusion from these studies is that the effectiveness of a centralized regime stands and falls with the ability and willingness of just one person to invest resources for punishment, which results in considerably more variability in performance between groups. Decentralized regimes, on the other hand, suffer from the problems already mentioned. Such negative effects are likely to be aggravated by noisy information, which so far was only tested in decentralized regimes. Several studies show that peer to peer punishment is not able to sustain cooperation under imperfect information (Grechenig et al., 2010; Ambrus and Greiner, 2012; Grechenig et al., 2015) and it remains unclear whether this equally holds for centralized punishment. For example, imperfect information may result in some to stop punishing altogether, even irrespective of total group contributions. While in a decentralized regime this could partly be compensated by more punishment of others, if punishment is centralized, this is not possible.

In order to test the effect of centralization *per se* under different degrees of *imperfect information*, we hold all other considerations constant across institutions: (1), (1) punishers cannot commit to punishment *ex ante*, (2) contributors cannot influence who is allowed to punish, (3) the direct consequences from punishment are the same in both institutions, and (4) there are no differences in externalities from punishment. We introduce noise in the signals about individual contributions in order to test whether decentralized or centralized punishment is more robust to imperfect information. By abstracting from

institutional factors, we return to the origins of formal punishment as a centralization of informal sanctioning regimes (Turnbull, 1962; Guala, 2012). Furthermore, by allowing participants to interact over 30 periods, we obtain enough observations for a detailed analysis of individual behavior and group dynamics.

To the best of our knowledge, we are the first to analyze the effects of centralization of punishment *per se* under *imperfect information*.

## 2. EXPERIMENTAL DESIGN

We use a standard finitely repeated linear public-good game with a voluntary contribution mechanism. Participants interact in groups of five over 30 periods in a partner design, where every period has two stages, a contribution and a punishment stage. In our set of experiments we have two treatments. In the first we compare two different punishment *institutions*: Decentralized (DEC) and Centralized punishment (CEN). In the second we control for the accuracy of the signal about the contributions in the group. More specifically, we compare a perfect signal with three different levels of noise, indicated by the probability  $\lambda$  of an accurate signal with  $\lambda = 1$ ,  $\lambda = 0.75$ , or  $\lambda = 0.50$ . Thus, the signal is either always accurate, or only 75 or 50% of the time. In the following we identify a treatment condition by the combination of institution and  $\lambda$ . For example DEC/1 is the treatment with decentralized punishment and a perfect signal.

There are five participants in every group, four *C-participants* ( $i \in \{1 \dots 4\}$ ) and one *Authority* (A), and every period consists of two stages, a contribution and a punishment stage. In the first stage, the contribution stage, the four *C-participants* can contribute to the public good; the remaining participant, the *Authority* (A), benefits from the public good but cannot contribute herself. After the contribution stage, all five participants first receive a common signal about the contributions, where the quality of the signal depends on  $\lambda$ . Then, in treatments with centralized punishment (CEN) the authority decides over punishment. In treatments with decentralized punishment (DEC) this is done by the *C-participants* and the Authority is merely passive. More specifically, she cannot influence punishment but is, nevertheless, affected by the contribution and punishment decisions of the four *C-participants*. In the following we describe each stage in more detail<sup>2</sup>.

### 2.1. Contribution Stage

In the first stage of each of the 30 rounds, each of the four *C-participants* receives an endowment of  $e_g = 20$  tokens. They then simultaneously and independently determine their contribution to the public good  $g_i$  with  $g_i \in G = \{0, 2, 4, \dots, 20\}$

In line with most experiments on decentralized punishment in public good games, we chose a marginal per capita return of 0.4. Hence, the monetary payoff of player  $i$  in the first stage is given by

<sup>1</sup> *Anti-social* punishment is directed at someone who contributed more than oneself (see e.g., Herrmann et al., 2008), *perverse* punishment at someone who contributed more than the group average in a symmetric setting (see e.g., Cinyabuguma et al., 2006).

<sup>2</sup> We will use the female genus for the authority and the male genus for the *C-participants*. The experiments, however, did not control for sex, and no loaded language was used.

$$\pi_i^1 = e_g - g_i + 0.4 \sum_k g_k \tag{1}$$

The authority *A*, despite not contributing, equally benefits from the public good:

$$\pi_A^1 = 0.4 \sum_k g_k \tag{2}$$

## 2.2. Punishment Stage

In the second stage, every C-participant *i* and authority *A* receives the same signal  $s_k$  about the contributions  $g_k$  of participant *k* with  $k \neq i$ . The signal itself is independently determined for every contribution in the group. With probability  $\lambda$  the signal is correct and  $s_k = g_k$ , with probability  $1 - \lambda$  the signal is incorrect. An incorrect signal is taken from the uniform distribution of all possible contributions excluding the actual one. More specifically,

$$s_k = \begin{cases} g_k & \text{with probability} = \lambda \\ \tilde{g}_k & \text{with probability} = 1 - \lambda \end{cases},$$

where  $\lambda$  is the accuracy of the signal and  $\tilde{g}_k$  is an independent realization out of the uniform distribution  $G \setminus \{g_k\}$ . As all participants receive the same signals about contributions of others, information conditions for punishers are constant across treatments. Note, however, that every C-participant always knows for sure how much he contributed himself but they do not know which signal others receive about their own contribution<sup>3</sup>.

C-participants were identified by numbers 1–4 which were assigned randomly anew every period and made identification of group members difficult. We contrast a perfect signal ( $\lambda = 1$ ) with three levels of noise:  $\lambda = 0.5$  and  $\lambda = 0.75$ . For example, for  $\lambda = 0.5$ , a contribution of 6 would lead to the signal “6” with probability 0.5, and to any other signal, each with probability 0.05 (0.5/10). Thus, despite considerable noise, a signal is still informative.

### 2.2.1. Punishment

In both punishment institutions, each of the four C-participants *i* receives a punishment endowment of  $e_p = 10$  punishment points, and authority *A* receives an additional endowment of  $e_p^A = 40$ . Otherwise the rules differ.

### 2.2.2. Punishment in DEC

In *DEC* the four C-participants can distribute punishment points amongst each other, where each point costs them one unit and also reduces the authority’s income by one<sup>4</sup>. Every received punishment point reduces the target participant’s income by three units. More specifically, denoting a punishment point sent by *i* to *j* with  $p_{ij}$ , the total payoff of participant *i* is:

$$\pi_i = \pi_i^1 + e_p - \sum_j p_{ij} - 3 \sum_j p_{ji}, \tag{3}$$

<sup>3</sup> Note that due to external validity we decided to keep this informational asymmetry which is a natural feature of these institutions.

<sup>4</sup> The authority’s income is reduced by one to guarantee symmetry with treatment *CEN*.

and the payoff of participant *A* is:

$$\pi_A = \pi_A^1 + e_p^A - \sum_i \sum_{j \neq i} p_{ij} \tag{4}$$

We include the authority *A* in treatment *DEC* as a passive participant in order to hold considerations, such as the externalities from punishment (see Engel and Rockenbach, 2009), constant across treatments.

### 2.2.3. Punishment in CEN

In *CEN*, only authority *A* can distribute punishment points. Every punishment point distributed by *A* reduces *A*’s payoff by one, the punished subject’s payoff by three, and the payoff of each other participant by 1/3. This keeps the overall costs of punishment constant across treatments (participants finance the punishment applied by *A*, except one’s own punishment). Furthermore, we kept the maximal amount of punishment points that could be assigned to one participant constant across treatments by limiting *A* to distributing at most 30 points to one individual. Thus, in *CEN*, final payoffs are determined as follows:

$$\pi_i = \pi_i^1 + e_p - \frac{1}{3} \sum_{j \neq i} p_{Aj} - 3p_{Ai}, \tag{5}$$

and

$$\pi_A = \pi_A^1 + e_p^A - \sum_j p_{Aj}, \tag{6}$$

where equivalently to  $p_{ij}$  we denote with  $p_{Aj}$  the number of punishment points assigned to *j* by *A*. This payoff structure keeps everything constant across institutions, including externalities, efficiency concerns, commitment, and punishment power, such that we exclusively explore the effect of centralization *per se*.

### 2.2.4. Information at the End of One Round

At the end of a round, every participant is informed about his income from the public good, the amount of punishment received, and his final total payoff. Thus, any inaccurate perception about one’s own income, resulting from noisy signals about group contributions, are resolved at the end of every round.

## 2.3. Equilibrium

To find the set of equilibria under the assumptions of selfishness, common knowledge of rationality and sequential rationality, we first look at the stage game. As punishment is costly, it is strictly dominated, both for *A* in *CEN*, as well as for the regular participants in *DEC*. Thus, play at the contribution stage remains unaffected and zero contribution by all remains the unique equilibrium in dominant strategies. As zero contribution and no punishment guarantees the minmax payoff, the equilibrium is not affected by the finite number of repetitions, and there should be no cooperation and no punishment throughout the 30 periods. This rationale applies to all treatments, such that there should be no differences between different punishment institutions and different levels of noise.

## 2.4. Setup

We use a  $2 \times 3$  factorial design between subjects, i.e., every subject participates in only one of our six treatment combinations. Subjects interact repeatedly over 30 periods in a partners design, i.e., groups are kept constant. All rounds are paid.

The experiments were run in the experimental laboratory of the University of Bonn (EconLab) and was programmed and conducted with the software z-Tree (Fischbacher, 2007). We ran a total of 12 sessions with 240 (20 per session) student participants (mostly undergraduate from various fields of study) divided into 48 groups (8 groups of 5 participants per treatment), and no participant took part in more than one session. We relied on ORSEE (Greiner, 2015) for recruitment.

For 100 experimental *Taler* participants earned €1. Sessions lasted for about 90 min (including admission and payment) and participants earned on average €14.28 (including a show up fee of €2.50), about USD 18.80, which is more than the usual hourly wage for student jobs.

## 3. RESULTS

We first report aggregate results with respect to institutions and noise before we analyze punishment behavior, the resulting incentives, and intertemporal reactions in more detail.

### 3.1. Aggregate Outcomes

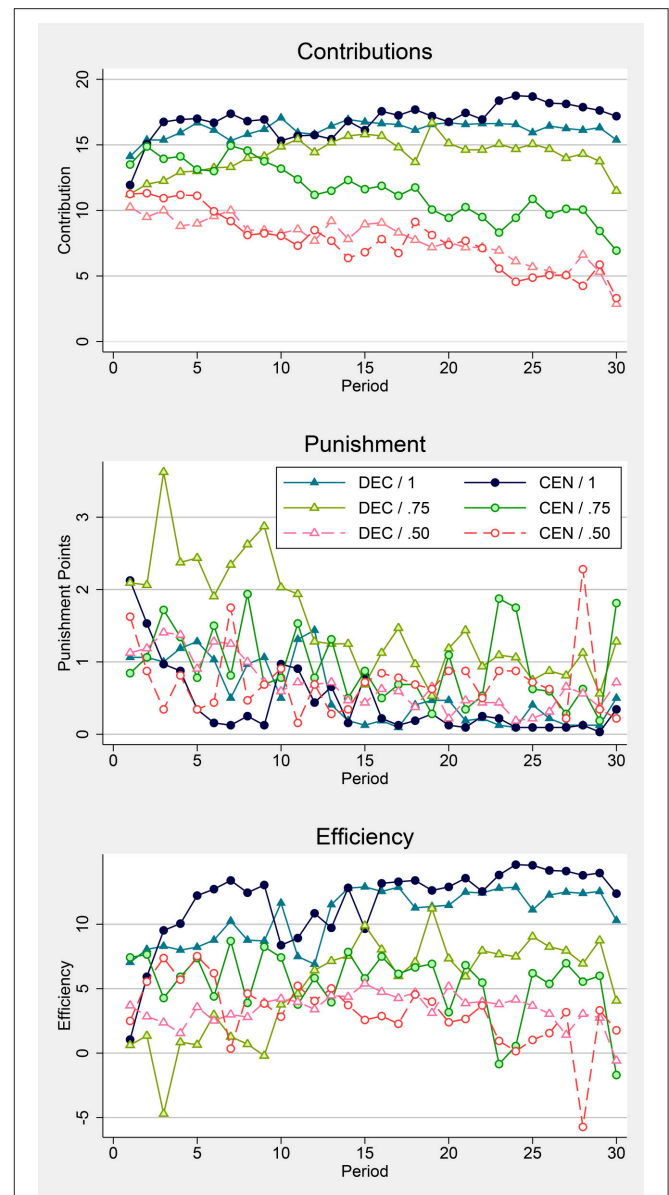
We compare contributions, overall punishment, and efficiency between institutions and noise levels. The graphs in **Figure 1** show average contributions, punishment points (distributed to C-participants), and efficiency over time across treatment combinations. Efficiency is measured as the difference between total earnings minus fixed endowments (for contribution and punishment stage), including authorities. Thus, an efficiency of 0 means that inefficiencies due to punishment balance out gains from cooperation.

#### 3.1.1. Differences between Institutions

**Table 1** reports the  $p$ -values of two-sided Wilcoxon rank-sum tests comparing the distributions of matching group averages between DEC and CEN within each noise condition (see rows)<sup>5</sup>. For each variable of interest, contributions, punishment points, and total efficiency, we test separately for the first (Periods 1–15) and second half (Periods 16–30) of the experiment (see columns) and rely on acceptance thresholds of 10%.

For a perfect signal, i.e., whenever  $\lambda = 1$  there are neither significant differences in contributions nor in punishment or efficiency. Similarly, for considerable noise of  $\lambda = 0.50$  **Table 1** does not report any significant differences between institutions. For  $\lambda = 0.75$ , however, we find significant differences between institutions. In the second half of the experiment contributions are significantly higher in DEC/0.75 than in CEN/0.75. Furthermore, there is significantly more punishment in DEC than in CEN in the first half of the experiment, but not in the second half. The substantially higher punishment in DEC/0.75 in the first half, and here especially in the first ten

<sup>5</sup> More specifically, we take the average for every matching group and compare the resulting distributions.



**FIGURE 1 | Outcomes.** Treatments with DECentralized or CENTralized punishment and a probability of a correct signal about individual contributions of 1, 0.75, or 0.50.

periods, results in significantly lower efficiency than in CEN/0.75. Despite contributions being higher in late rounds, there are no significant differences in efficiency. If instead of total efficiency we only look at *Profits* of C-participants, results are very similar and qualitatively identical for all levels of  $\lambda$ .

**Result 1.** *If signals are perfect ( $\lambda = 1$ ) or highly distorted ( $\lambda = 0.5$ ) aggregate outcomes do not differ significantly between institutions. For  $\lambda = 0.75$  there is significantly more punishment and less efficiency in DEC early in the experiment, and in later periods contributions remain significantly higher in DEC/0.75 than in CEN/0.75.*

TABLE 1 | Aggregate effects of institution.

Periods:	Contrib.		Punishment		Efficiency	
	1–15	16–30	1–15	16–30	1–15	16–30
$\lambda = 1$	0.752	0.709	0.713	0.594	0.834	0.709
$\lambda = 0.75$	0.753	<b>0.046</b>	<b>0.027</b>	0.462	0.142	0.142
$\lambda = 0.50$	0.495	0.674	0.793	0.400	0.753	0.529

Each cell reports the  $p$ -value of a Wilcoxon rank-sum test comparing the distributions of averages per matching group between CEN and DEC for a given  $\lambda$  (row). Averages per matching group are either taken over first or last 15 periods. Significant outcomes with  $p < 0.1$  are highlighted.

Most contributions in CEN/1 and DEC/1 are close to the ceiling of 20. In Section A.2. (Supplementary Material) of the online supplement we report simulations which show that despite of this we could have observed significant effects, if in DEC/1 a sufficient number of participants had contributed fully by default. For example, suppose that participants in DEC/1 had contributed fully with a baseline chance of 70 or 80%. In this case we could have found a significant difference between DEC/1 and CEN/1 at the 5% significance level with a probability of 52.3 and 94.6%, respectively.

### 3.1.2. Effect of Noise

Comparisons between different levels of noise within each institution mirror our previous results. Table 2 lists the  $p$ -values of Wilcoxon rank sum tests comparing outcomes between different levels of noise within each institution. Again we distinguish between the first and second half of the experiment (see columns).

In CEN contributions are always significantly smaller the more noisy the signal. In DEC, however, there are no significant differences between  $\lambda = 1$  and  $\lambda = 0.75$  in the first half of the experiment, while otherwise comparative statics are as in CEN. Aggregate Punishment, on the other hand remains unaffected by noise in CEN in the first half, but not in the second, where it is smaller in CEN/1 than in CEN/0.75 and CEN/0.50. In DEC only  $\lambda = 0.75$  stands out. While punishment never differs between DEC/1 and DEC/0.50, there is significantly more punishment in DEC/0.75 than in either DEC/1 or DEC/0.50 in the first half and significantly more than in DEC/1 also in the second half. In both institutions, efficiency is similarly affected by noise in the first half of the experiment. It is significantly higher for  $\lambda = 1$  than for  $\lambda = 0.75$  or  $\lambda = 0.50$ . While there are no significant differences between CEN/0.75 and CEN/0.50 throughout the experiment, efficiency is significantly higher for DEC/0.75 than for DEC/0.50 in the second half, a consequence of the higher contributions towards the end of DEC/0.75. In summary we find an interaction effect between institution and noise:

**Result 2.** Irrespective of institution, noise has a negative effect on contributions in the long run. The effect of noise on punishment and efficiency, however, is non-monotonic and differs between institutions. Under centralized punishment CEN/1 stands out with significantly less punishment and higher efficiency in the long run,

TABLE 2 | Aggregate effects of noise.

Periods:	DEC		CEN	
	1–15	16–30	1–15	16–30
<b>Contributions</b>				
$\lambda = 1$ vs. 0.75	0.208	<b>0.059</b>	<b>0.092</b>	<b>0.010</b>
$\lambda = 1$ vs. 0.50	<b>0.009</b>	<b>0.012</b>	<b>0.021</b>	<b>0.002</b>
$\lambda = 0.75$ vs. 0.5	<b>0.059</b>	<b>0.002</b>	<b>0.059</b>	<b>0.074</b>
<b>Punishment</b>				
$\lambda = 1$ vs. 0.75	<b>0.046</b>	<b>0.016</b>	0.401	<b>0.026</b>
$\lambda = 1$ vs. 0.50	1.00	0.342	0.916	<b>0.023</b>
$\lambda = 0.75$ vs. 0.5	<b>0.021</b>	0.103	0.431	0.916
<b>Efficiency</b>				
$\lambda = 1$ vs. 0.75	<b>0.059</b>	<b>0.074</b>	<b>0.074</b>	<b>0.003</b>
$\lambda = 1$ vs. 0.50	<b>0.059</b>	<b>0.021</b>	<b>0.016</b>	<b>0.002</b>
$\lambda = 0.75$ vs. 0.5	0.600	<b>0.005</b>	0.401	0.294

Each cell reports the  $p$ -value of a Wilcoxon rank-sum test comparing distributions of averages per matching group between different levels of noise within an institution. Averages per matching group are either taken over first or last 15 periods. Significant outcomes with  $p < 0.1$  are in bold.

while there are hardly any differences between CEN/0.75 and CEN/0.50. Under decentralized punishment, DEC/0.75 stands out with more punishment and less efficiency especially in early rounds.

## 3.2. Distributed Punishment

We turn to regression analysis to identify how the decision to punish is correlated to signals about contributions. The columns in Table 3 report the results of estimations of the distributed amount of punishment points on characteristics of the received signals, separately for every treatment<sup>6</sup>. Variable  $D_{Free}$  is a dummy variable indicating a signal of no contribution. Variables  $D_{aboveGroup}$  and  $D_{belowGroup}$  are dummy variables indicating that, according to the signal, the receiver of punishment has contributed more than the group average and less, respectively. In CEN we calculate the average group contribution as the average signal. In DEC we take the average over one's own contribution and the signals about the contributions of the other three. Similarly, variables  $D_{more}$  and  $D_{less}$  in treatment DEC are dummy variables indicating that, according to the signal, the (potential) receiver of punishment contributed more than the one who decides whether to punish and less, respectively. The definitions of  $D_{more}$ ,  $D_{less}$ ,  $D_{aboveGroup}$ , and  $D_{belowGroup}$  all exclude the case of no difference (avoiding perfect collinearity)<sup>7</sup>. Our measurement of cooperativeness as the deviation to the group average is similar to, for example, (Cinyabuguma et al., 2006). The measures based on

<sup>6</sup>Whenever we speak of a "signal," we are referring to the number the punisher sees on her screen about the contribution of the other participant.

<sup>7</sup> Thus, in CEN the constant of the regression measures average punishment of someone who contributed exactly as much as the group average. Similarly, in DEC it measures the punishment of someone whose contribution does not only match the group average but also equals the contribution of the participant who decides whether to punish.

TABLE 3 | Distributed punishment.

	$\lambda = 1$		$\lambda = 0.75$		$\lambda = 0.50$	
	DEC	CEN	DEC	CEN	DEC	CEN
<b>DISTRIBUTED PUNISHMENT POINTS</b>						
Signal	-0.0333 (-0.48)	» -0.127*** (-11.73)	-0.0731*** (-3.36)	» -0.105*** (-6.13)	-0.134*** (-4.83)	« -0.0511 (-0.72)
DFreerider	0.792** (2.54)	» -0.110 (-0.29)	0.0223 (0.09)	< 0.256 (0.73)	0.134 (0.78)	< 0.494 (1.41)
DbelowGroup	2.141*** (3.12)	« 3.104*** (5.13)	1.191*** (5.68)	1.029 (1.30)	0.699* (1.96)	< 1.400 (1.38)
daboveGroup	1.653*** (2.89)	> 0.503 (0.62)	0.168 (0.71)	0.117 (0.17)	0.496** (2.06)	< 1.468 (1.52)
Dless	2.056*** (5.39)		1.792*** (8.63)		0.606*** (2.92)	
Dmore	1.689*** (4.01)		1.297*** (6.96)		0.737*** (5.92)	
Period	0.138 (0.43)	< 0.478*** (3.21)	0.281 (1.28)	< 0.437* (1.85)	-0.206 (-0.88)	» -0.772 (-1.43)
_cons	-5.080*** (-8.93)	« -1.843** (-2.16)	-3.316*** (-5.15)	« -1.116 (-0.96)	-2.643*** (-3.55)	-0.260 (-0.15)
N (#Groups)	2880 (8)	960 (8)	2880 (8)	960 (8)	2880 (8)	960 (8)
ll	-935.8	-393.7	-1940.2	-1137.1	-1141.8	-1078.8
Wald chi2	645.0***	954.5***	1333.3***	946.0***	513.5***	385.0***

Poisson regression of distributed punishment points with random shift effect on matching group and subject (DEC only, random effects independent). Not reported: fixed effect on every Period (following Period 2). *t* statistics in parentheses based on Huber White standard errors (not for Wald chi2). \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ . Significance of differences between institutions (based on regression of data on both institutions, including dummy for treatment CEN and interaction effects): < ( $p < 0.1$ ), < ( $p < 0.05$ ), « ( $p < 0.01$ ).

individual comparisons in DEC are as in Herrmann et al. (2008). Clearly, both measures are related<sup>8</sup>.

All regressions are maximum likelihood Poisson estimations which besides the above mentioned fixed effects have a random intercept effect per matching group, and in DEC also per participant (independent of group effect). Estimations also include fixed effects for periods (following period 2). We use Poisson estimations as this best controls for the skewed distribution of integer punishment points.

While in DEC patterns of punishment decisions are similar for all levels of  $\lambda$ , in CEN punishment becomes fairly undirected with increasing noise. In all treatment combinations, punishment decreases with increasing Signal about the contribution. Only in DEC/1 and CEN/0.50 this is not significant. Of all treatment combinations, only DEC/1 has significant additional punishment of free rider behavior (DFreerider). In all treatments someone gets punished more when the signal about his contribution is below the group average (DbelowGroup), an effect which is significant in all treatments except CEN/0.75 and CEN/0.50. Similarly, participants who contribute less than the punisher (Dless) in DEC are punished significantly more.

In DEC/1 and DEC/0.50 we furthermore find significant perverse punishment as measured by DaboveGroup, an effect

<sup>8</sup> Furthermore, as Steiner (2007) shows, a punishment rule exclusively targeting the lowest contributor can establish cooperation even if punishment sentiments are very small. It is therefore reasonable to measure cooperativeness in relation to average group contributions.

which is absent in all CEN treatments. Furthermore, in all DEC treatments there is significant anti-social punishment as indicated by Dmore.

### Result 3. Distributed punishment

- becomes less systematic with increasing noise in CEN,
- and decreases with increasing signal in all treatments.
- There is perverse punishment in DEC/1 and DEC/0.5 but in none of the CEN treatments.
- There is anti-social punishment in DEC for all levels of  $\lambda$ .

We furthermore tested whether there are different patterns in punishment behavior between the first 10, 15, last 15, and last 20 periods, especially in DEC/0.75. While there are some minor differences in the size of reactions, the overall patterns are similar. We furthermore tested whether differences in coefficients between, for example DEC/1 and CEN/1 are significant<sup>9</sup>. Significant differences are indicated by symbols such as > ( $p < 0.1$ ), > ( $p < 0.05$ ), or » ( $p < 0.01$ ) between the two columns.

### Result 4. With only a few exceptions, punishment behavior differs significantly between institutions.

<sup>9</sup> For this purpose we introduced a dummy variable for institution DEC and interacted it with all variables before running a regression on the combined data for both institutions within one noise treatment.

TABLE 4 | Received punishment.

	$\lambda = 1$		$\lambda = 0.75$		$\lambda = 0.50$	
	DEC	CEN	DEC	CEN	DEC	CEN
<b>DEDUCTED INCOME</b>						
Contribution	-0.0385 (-0.61)	>> -0.140*** (-5.71)	-0.0360** (-2.28)	<< 0.00296 (0.17)	-0.0616*** (-5.21)	<< -0.0193 (-0.63)
DFreerider	0.904*** (2.80)	>> 0.139 (0.33)	-0.117 (-0.46)	<< 1.083*** (5.73)	0.281* (1.70)	0.446 (1.32)
DbelowGroup	3.511*** (5.98)	>> 2.871*** (4.58)	0.696** (2.28)	0.597** (2.23)	-0.196 (-0.22)	< 0.408** (2.45)
DaboveGroup	2.852*** (7.02)	>> 0.604 (0.70)	-0.103 (-0.42)	-0.111 (-0.49)	0.0945 (0.12)	0.268 (0.98)
Period	0.035 (0.62)	>> 0.00272 (0.19)	0.0106 (0.53)	-0.0171 (-0.40)	-0.067 (-1.64)	>> -0.116*** (-3.03)
_cons	-2.172** (-2.29)	<< -0.023 (-0.02)	1.734*** (4.96)	>> 0.045 (0.09)	1.01 (0.87)	0.727 (1.22)
N (#Groups)	960 (8)	960 (8)	960 (8)	960 (8)	960 (8)	960 (8)
Wald chi2	1325.2***	2154.5***	1594.7***	1634.1***	607.7***	787.8***
ll	-1381.3	-806.9	-4785.8	-3521.4	-2090.1	-2789.4

Poisson regression of deducted income with random shift effect on subject (DEC only) and matching group (independent). Not reported: fixed effect on every Period (following Period 2). *t* statistics in parentheses based on Huber White standard errors (not for Wald chi2). \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ . Significance of differences between institutions (based on regression of data on both institutions, including dummy for treatment CEN and interaction effects): < ( $p < 0.1$ ), < ( $p < 0.05$ ), << ( $p < 0.01$ ).

Finally, which effect is stronger, punishment of negative or positive deviators?<sup>10</sup> With the exception of DEC/0.50, punishment of negative deviators is stronger. Wald tests comparing coefficients *D*below with *D*above or *D*less with *D*more confirm that these differences are significant in treatments with  $\lambda = 1$  or  $\lambda = 0.75$  (all  $p$ -values below 0.0451), but not for  $\lambda = 0.50$  ( $p \geq 0.2073$ )

Result 5. *Except for treatment combinations with  $\lambda = 0.50$ , punishment of negative deviators is significantly stronger than that of positive deviators.*

### 3.3. Incentives

To adequately compare institutions we must compare how punishment is being experienced. In treatment DEC recipients of punishment can not identify individual punishers and therefore whether the punisher contributed more or less. Thus, contrary to distributed punishment, *D*more and *D*less are now undetermined in both institutions, which puts the two back on equal footing. Furthermore, for DEC it is conceivable that due to the many repetitions, behavior of “sophisticated” punishers may add a non-random component to how distributed punishment is aggregated in the group<sup>11</sup>. So, while in CEN received punishment must be equivalent to the way punishment is distributed, in DEC received punishment may be more than just a combination of distributed punishment with white noise.

<sup>10</sup> We thank an anonymous referee for suggesting this analysis.

<sup>11</sup> Although, very difficult to achieve, there could be some form of “tacit” collusion amongst punishers, and participants may learn to identify the accuracy of signals, for example based on repetitive or unique contribution behavior.

Table 4 reports results of Poisson regressions of the received punishment on various characteristics of a participant's contribution. All regressors are now based on actual contributions rather than signals. This includes the definition of the group average for the calculation of *D*aboveGroup and *D*belowGroup.

For perfect signals our estimations by and large confirm our previous results<sup>12</sup>. Again we find significant punishment of negative deviators in both CEN/1 and DEC/1, and of positive deviators (*D*aboveGroup) in DEC/1 only. Contrary to our previous results, neither in DEC/0.75 nor in DEC/0.50 do participants experience significant perverse punishment. Despite the noise, in all treatments except for DEC/0.50, negative deviators receive significantly more punishment. Finally, there is no experience of perverse punishment in any of the CEN treatments.

Result 6. • *In DEC/1 negative and positive deviators experience punishment.*

- *In DEC/0.75 only negative deviators experience punishment.*
- *In DEC/0.50 received punishment is independent of how one's contribution relates to those in the group.*
- *In all CEN treatments negative deviators experience significantly more punishment while positive deviators do not.*

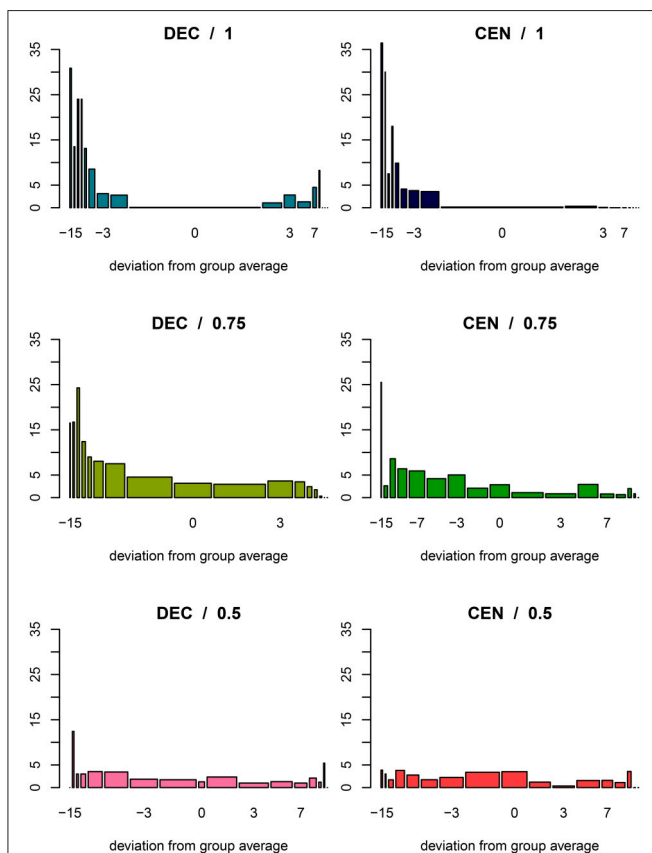
Again we compared coefficients between institutions. Significant results are indicated by symbols such as > between columns in

<sup>12</sup>It might be surprising that coefficients for CEN/1 in Tables 3, 4 differ. However, in the former, observations are clustered on the authority, compared to the recipient of punishment in the latter.

**Table 4.** Similarly to our results on distributed punishment we find:

Result 7. *With only a few exceptions, experienced punishment differs significantly between institutions.*

In addition to the regressions reported in **Table 4** we ran estimations including variables *above* and *below* which measure the absolute distance of the contribution to the group average (rather than merely indicator variables *DaboveGroup* and *DbelowGroup*). We report detailed results in the online supplement. The barplots in **Figure 2** illustrate the average amount of punishment participants received conditioned on how much their contribution deviated from the group average. The height of the bars indicate the average punishment, and the width is proportional to the number of observations in that category. Our main finding is that only in DEC/1, DEC/0.75, and CEN/1 the size of punishment is overall positively correlated to the absolute size of the deviation from the group average. There is, however, one important difference between treatments DEC and CEN. In the former this overall correlation is driven by a strong partial effect on the deviation, whereas in CEN it is driven by a strong reaction to the absolute size of the contribution.



**FIGURE 2 | Received punishment.** Barplot of average deducted income by difference between own contribution and average group contribution. Except for 0, categories show intervals of length 2: “-16 to -14”, “-14 to 12” ... “-2 to < 0”, “0”, “> 0-2”, ... “14-16”. Width of bars is proportional to number of observations.

### 3.4. Reaction to Punishment

**Table 5** reports the result of three linear regressions of changes in contributions ( $\text{contribution}_t - \text{contribution}_{t-1}$ ) on outcomes in the previous round - one regression each for every level of noise. More specifically, the first estimation (first two columns) includes data from  $\lambda = 1$  treatments only, and we estimated the average change in contributions for DEC and CEN for six possible outcomes in the previous round. The outcomes in the previous round are identified by (i) whether the participant contributed more than the group average (*above*), the same amount (*equal*), or less (*below*), and (ii) whether she was punished or not. Every regressor is a dummy variable indicating the relevant combination and all calculations are based on actual contributions rather than signals. The maximum likelihood estimations correct for random effects on participants and group<sup>13</sup>. For example, in DEC/1 someone who in the previous period contributed more than the group mean, on average reduces the contribution by 1.025 if there was no punishment and by 1.505 if there was. The table also reports results from one sided Wald tests comparing reactions between institutions (DEC vs. CEN, significance indicated, e.g., by > or <<) as well as testing for a significant positive effect of punishment on reactions (for example, define  $\delta = \text{“below \& punished”} - \text{“below \& not pun,”}$  then we test  $H_0: \delta = 0$  vs.  $H_1: \delta > 0$ . Significance is indicated by, e.g.,  $\Delta$ ).

In both institutions and for all levels of noise, those who in the previous round contributed less, on average significantly increase their contribution, irrespective of whether they were punished or not. In all treatments except CEN/0.5, punishment significantly enhances this reaction. Those who contributed more and were not punished decrease their contributions in all treatment combinations. In DEC, punishment of such cooperative participants has no significant effect on this reaction. In CEN, however, such perverse punishment dampens the reduction, significantly so in CEN/1 and CEN/0.75. Comparing immediate reactions between institutions, treatments CEN/1 and CEN/0.75 in most cases show significantly more cooperative changes in contributions than the equivalent DEC treatments. Note, however, that there are other reactions that span over more periods.

In Table A.5 (Supplementary Material) of the online supplement we report results from similar regressions including the amount of the received punishment for each of the three possible cases (*above* or *below* average or *equal* contribution). Including the amount of the received punishment into the regression increases the shift effect (or “intercept”) throughout, and with only one exception (equal contributors in DEC/1), the reaction to the amount of punishment is negative. For those contributing below average it is significantly negative in all treatments except for CEN/0.75, and for above average contributors it is significantly negative in all except for DEC/0.50 and CEN/0.50.

<sup>13</sup> More specifically, reported reactions are the fixed effects from linear maximum likelihood estimations including independent random effects on participants nested in random effects on matching groups. As in **Table 4**, variables are defined by actual contributions rather than signals.



TABLE 5 | Change in contributions.

	$\lambda = 1$		$\lambda = 0.75$		$\lambda = 0.50$	
	DEC	CEN	DEC	CEN	DEC	CEN
<b>CHANGE IN CONTRIBUTION</b>						
<i>D</i> below & not pun.	1.146** (2.15)	< 2.630*** (5.03)	1.220** (2.41)	< 2.734*** (6.26)	1.966*** (4.61)	2.623*** (6.25)
	$\Delta$	$\Delta$	$\Delta$	$\Delta$	$\wedge$	
<i>D</i> below & punished	3.681*** (9.45)	< 4.968*** (12.31)	3.260*** (7.54)	$\ll$ 4.956*** (10.22)	3.067*** (6.01)	2.587*** (4.89)
<i>D</i> equal & not pun.	-0.394* (-1.68)	-0.498** (-2.05)	-0.711 (-1.29)	-0.54 (-0.79)	-0.32 (-0.28)	1.063 (1.45)
	$\wedge$		$\wedge$	$\wedge$		$\wedge$
<i>D</i> equal & punished	0.9 (0.93)	> -0.853 (-1.01)	-0.947 (-1.02)	-2.699** (-2.02)	3.053 (1.02)	2.825*** (3.34)
<i>D</i> above & not pun.	-1.025*** (-2.72)	-1.266*** (-4.22)	-1.900*** (-4.77)	$\gg$ -3.461*** (-9.8)	-3.286*** (-8.07)	-3.937*** (-9.62)
		$\Delta$		$\Delta$	$\wedge$	
<i>D</i> above & punished	-1.505*** (-3.19)	$\ll$ 1.812 (1.61)	-1.734*** (-3.4)	-1.714** (-2.39)	-2.095*** (-3.46)	> -3.943*** (-5.54)
N/Subj./Groups	1856/64/16		1856/64/16		1856/64/16	
chi2	397.1***		438.7***		448.9***	
ll	-5053.7		-5716.8		-5702.7	

Fixed effects margins (t-statistics) from linear maximum likelihood regression with random effect on participant nested in (independent) random effect on matching group. \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ . Significance of differences between institutions (based on Wald test, one-sided): < ( $p < 0.1$ ), < ( $p < 0.05$ ),  $\ll$  ( $p < 0.01$ ). Significance of effect of punishment (based on Wald test, one-sided):  $\wedge$  ( $p < 0.1$ ),  $\wedge$  ( $p < 0.05$ ),  $\Delta$  ( $p < 0.01$ ).

#### Result 8. For negative deviators we find:

- Negative deviators increase their contribution in the next period in both institutions.
- This effect is stronger if they were punished and punishment was not too high.
- A large amount of punishment has a negative effect on contributions.
- These immediate reactions are stronger in CEN than in DEC.

#### For positive deviators we find:

- Positive deviators decrease their contribution in the next period.
- In DEC punishment has no effect.
- In CEN punishment tends to reduce this adjustment.
- However, a lot of punishment has a negative effect on contributions in DEC and CEN.

We furthermore tested whether reactions to punishment differ in early rounds. By and large results are similar and do not differ significantly from the one we reported. The most noteworthy (though insignificant) difference is that in CEN/1 in early rounds, the positive reaction of cooperative types to punishment is much smaller and only later becomes significant.

Finally, to fully assess the question whether punishment has a stronger effect on contribution behavior in DEC or CEN, we compare the change in contribution due to punishment

on top of the change in contribution without punishment<sup>14</sup>. More specifically, in both institutions deviators change their contribution even if they are not punished. The effectiveness of punishment is therefore the difference in adjustments. For example, define by  $\Delta b$  (DEC/1) the following differences in coefficients from Table 5.  $\Delta b = (\text{below \& punished}) - (\text{below \& not pun.})$ . We compared these differences between institutions by testing for example  $H_0 : \Delta b \text{ (DEC/1)} = \Delta b \text{ (CEN/1)}$  against  $H_1 : \Delta b \text{ (DEC/1)} > \Delta b \text{ (CEN/1)}$ . The result of these tests are indicated by appropriate signs between the columns of Table 6.

Result 9. For  $\lambda = 1$  and  $\lambda = 0.75$ , the effectiveness of punishment on contributions only differs significantly between institutions for positive deviators. For  $\lambda = 0.50$ , it differs significantly for negative deviators.

## 4. DISCUSSION

We test whether centralization of punishment *per se* under *perfect* and *imperfect* information affects behavior and outcomes in voluntary contribution public good games. Both, institution and noisiness of signals are imposed exogenously, and efficiency and cost of punishment are held constant across institutions.

We find significant differences in how contribution behavior is being punished. While centralized authorities do not

<sup>14</sup> We thank an anonymous referee for suggesting this analysis.

**TABLE 6 | Difference in change in contributions after punishment vs. no punishment.**

	$\lambda = 1$		$\lambda = 0.75$		$\lambda = 0.50$	
	DEC	CEN	DEC	CEN	DEC	CEN
$\Delta_{\text{below}}$	2.535***	2.338***	2.040***	2.222***	1.101**	> -0.036
$\Delta_{\text{equal}}$	1.294*	-0.355	-0.236	-2.159*	3.373	1.762*
$\Delta_{\text{above}}$	-0.480	≪ 3.078***	0.166	< 1.747***	1.191**	-0.006

Coefficients and Wald-tests based on results from **Table 5**. For example:  $\Delta_{\text{below}} = (\text{below \& punished}) - (\text{below \& not pun.})$ . \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ . Significance of differences between institutions: < ( $p < 0.1$ ), << ( $p < 0.05$ ), ≪ ( $p < 0.01$ ).

punish those who contribute more than average, we observe such perverse and furthermore anti-social punishment in the decentralized regime. While in treatments with noise, this perverse punishment is being averaged out, it remains highly relevant if signals are perfect. Otherwise we observe socially reasonable punishment of those who contribute less than the group average in all treatments. Similarly, distributed and received punishment decrease with signal and actual contribution, respectively.

Reactions to punishment by negative deviators are similar between institutions. Even if they are not punished they increase their contribution but punishment enhances this positive adjustment in both institutions by about the same. Positive deviators on the other hand, reduce their contributions in both institutions as long as they were not punished. Interestingly, perverse punishment by their peers has no effect on this reaction in DEC. Thus, while perverse punishment directly induces inefficiencies, it has hardly any consequences in the long run. In CEN on the other hand, where this type of punishment was not as pronounced, it has significant positive effects on contributions.

Despite these differences in behavior, we find that the institution has no effect, neither on contributions, aggregate punishment nor overall efficiency, both if information is perfect ( $\lambda = 1$ ) and if there is considerable noise ( $\lambda = 0.5$ ). Under intermediate noise ( $\lambda = 0.75$ ), however, punishment differs significantly but with only limited consequences for contributions and overall efficiency. More specifically, in the peer to peer punishment institution, punishment escalates during the first third of the experiment keeping contributions higher than under centralized punishment. However, due to the excessive punishment, efficiency is not significantly different.

Anti-social and perverse punishment are frequently observed in public good experiments with peer to peer punishment. In a detailed study of such behavior (Herrmann et al., 2008) suggest dominance, competitive personality, the desire to maximize one's relative payoff, or normative conformity as possible motivating factors. However, there is no obvious reason why these motivating factors should not exist among authorities. The absence of perverse punishment in the CEN treatments therefore suggests that the direct involvement in the contribution stage plays an important role. Alternative explanations such as do-gooder derogation (Monin, 2007) may, however, still apply. Do-gooder derogation requires that one's own uncooperative behavior is illustrated and made worse by

the exemplary behavior of the cooperative players. As authorities cannot contribute, they are clearly unaffected by this. Other explanations referring to the repeated interactions, such as counter punishment to prevent future punishment, retaliation, or "feuds" among participants may also explain this asymmetry among institutions. The fact that with increasing noise such punishment behavior becomes less relevant (both in distributed and received punishment) supports this interpretation. The differences in reactions to perverse punishment in DEC and CEN are another piece in the puzzle which is probably due to ascribed motives. Centralized authorities mainly punish cooperators if they wish them to increase their contributions further, whereas under peer-to-peer punishment this is probably not the reason.

What remains difficult to explain is the escalation of punishment in DEC/0.75 which is absent in the equivalent treatment with centralized punishment CEN/0.75. Could it be explained by ruthless satisfaction of punishment sentiments and expression of anger (Dickinson and Masclet, 2015) despite the noise? But why would peers in a decentralized punishment mechanism succumb to such behavior while authorities do not? A possible explanation is that such sentiments are stronger among peers. However, this raises the question why we do not find differences between DEC/1 and CEN/1.

Centralized punishment institutions have been praised in the economic, legal, and political science literature, with recent support from experimental research. However, it is not centralization *per se* that is beneficial but other institutional differences. This issue carries over to studies with *endogenous* institutions, where centralized punishment prevails if it comes with additional advantages (Traulsen et al., 2012), but loses against decentralized punishment in a *ceteris paribus* comparison under perfect information (Grechenig et al., 2015).

## AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

## FUNDING

We gratefully acknowledge financial support from the Max Planck Institute for Research on Collective Goods, Bonn, Germany.

## ACKNOWLEDGMENTS

We thank Timothy Cason, Sebastian Goerg, and the three referees, one of them anonymous, for their helpful comments.

## REFERENCES

- Ambrus, A., and Greiner, B. (2012). Imperfect public monitoring with costly punishment: an experimental study. *Am. Econ. Rev.* 102, 3317–3332. doi: 10.1257/aer.102.7.3317
- Ambrus, A., and Greiner, B. (2015). Democratic punishment in public good games with perfect and imperfect observability. *Work. Pap.* doi: 10.2139/ssrn.2567326. [Epub ahead of print].
- Andreoni, J., and Gee, L. (2012). Gun for hire: does delegated enforcement crowd out peer punishment in giving to public goods? *J. Public Econ.* 96, 1036–1046. doi: 10.1016/j.jpubeco.2012.08.003
- Baldassarri, D., and Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1–5. doi: 10.1073/pnas.1105456108
- Carpenter, J., Kariv, S., and Schotter, A. (2012). Network architecture, cooperation and punishment in public good experiments. *Rev. Econ. Des.* 16, 93–118. doi: 10.1007/s10058-012-0120-z
- Cinyabuguma, M., Page, T., and Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Exp. Econ.* 9, 265–279. doi: 10.1007/s10683-006-9127-z
- Clotfelter, C. T. (1978). Private security and the public safety. *J. Urban Econ.* 5, 388–402. doi: 10.1016/0094-1190(78)90018-9
- Dickinson, D. L., and Masclot, D. (2015). Emotion venting and punishment in public good experiments. *J. Public Econ.* 122, 55–67. doi: 10.1016/j.jpubeco.2014.10.008
- Engel, C., and Rockenbach, B. (2009). *We Are Not Alone: The Impact of Externalities on Public Good Provision. Preprint 29/2009*. Bonn: Max Planck Institute for Research on Collective Goods.
- Fehr, E., and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* 90, 980–994. doi: 10.1257/aer.90.4.980
- Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10, 171–178. doi: 10.1007/s10683-006-9159-4
- Gächter, S., Renner, E., and Sefton, M. (2008). The long-run benefits of punishment. *Science* 322, 1510. doi: 10.1126/science.1164744
- Grechenig, K., Nicklisch, A., and Thöni, C. (2010). Punishment despite reasonable doubt - a public goods experiment with sanctions under uncertainty. *J. Emp. Legal Stud.* 7, 847–867. doi: 10.1111/j.1740-1461.2010.01197.x
- Grechenig, K., Nicklisch, A., and Thöni, C. (2015). *Information-Sensitive Leviathans: The Emergence of Centralized Punishment*. WiSo-HH Working Paper Series.
- Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *J. Econ. Sci. Assoc.* 1, 114–125. doi: 10.1007/s40881-015-0004-4
- Guala, F. (2012). Reciprocity: weak or strong? what punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* 35, 45–59. doi: 10.1017/S0140525X1100166X
- Guillen, P., Schwieren, C., and Staffiero, G. (2007). Why feed the leviathan? *Public Choice* 130, 115–128. doi: 10.1007/s11127-006-9075-3
- Herrmann, B., Thöni, C., and Gächter, S. (2008). Antisocial punishment across societies. *Science* 319, 1362–1367. doi: 10.1126/science.1153808
- Kosfeld, M., and Riedl, A. (2004). *The Design of (de)centralized Punishment Institutions for Sustaining Cooperation*. Tinbergen Institute Discussion Paper.
- Kube, S., and Traxler, C. (2011). The interaction of legal and social norm enforcement. *J. Public Econ. Theor.* 13, 639–660. doi: 10.1111/j.1467-9779.2011.01515.x

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fnbeh.2016.00180>

- Monin, B. (2007). Holier than me? threatening social comparison in the moral domain. *Int. Rev. Soc. Psychol.* 20, 53–68.
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Public Econ.* 92, 91–112. doi: 10.1016/j.jpubeco.2007.04.008
- Nikiforakis, N. (2014). “Self-Governance through altruistic punishment?” in *Reward and Punishment in Social Dilemmas*, eds P. A. M. Van Lange, B. Rockenbach, and T. Yamagishi (New York, NY: Oxford University Press), 197–213.
- Nikiforakis, N., Noussair, C., and Wilkening, T. (2012). Normative conflict and feuds: the limits of self-enforcement. *J. Public Econ.* 96, 797–807. doi: 10.1016/j.jpubeco.2012.05.014
- Nosenzo, D., and Sefton, M. (2014). “Promoting cooperation: the distribution of reward and punishment power,” in *Reward and Punishment in Social Dilemmas, Series in Human Cooperation*, eds P. A. Van Lange, B. Rockenbach, and T. Yamagishi (New York, NY: Oxford University Press), 87–114.
- O’Gorman, R., Henrich, J., and Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proc. Biol. Sci.* 276, 323–329. doi: 10.1098/rspb.2008.1082
- Ostrom, E., Walker, J., and Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *Am. Polit. Sci. Rev.* 86, 404–417. doi: 10.2307/1964229
- Polinsky, A. M. (1980). Private versus public enforcement of fines. *J. Legal Stud.* 9, 105–127. doi: 10.1086/467630
- Putterman, L., Tyran, J.-R., and Kamei, K. (2011). Public goods and voting on formal sanction schemes. *J. Public Econ.* 95, 1213–1222. doi: 10.1016/j.jpubeco.2011.05.001
- Steiner, J. (2007). A trace of anger is enough: on the enforcement of social norms. *Econ. Bull.* 8, 1–5.
- Traulsen, A., Röhl, T., and Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B* 279, 3716–3721. doi: 10.1098/rspb.2012.0937
- Turnbull, C. M. (1962). *The Forest People*. New York, NY: Touchstone.
- Tyran, J.-R., and Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *Scand. J. Econ.* 108, 135–156. doi: 10.1111/j.1467-9442.2006.00444.x
- Weber, M. (1919). “Politics as a vocation,” in *From Max Weber: Essays in Sociology/edited, with an introduction by H. H. Gerth and C. Wright Mills with a new preface by Bryan S. Turner 2009*, eds H. H. Gerth, and C. W. Mills (London; New York, NY: Routledge Taylor & Francis Group), 77–128.
- Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* 51, 110–116. doi: 10.1037/0022-3514.51.1.110

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer DN and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2016 Fischer, Grechenig and Meier. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.