Economics
Working Papers
2020-15

Cooperation and Norm-Enforcement under Impartial vs. Competitive Sanctions

Jan Philipp Krügel and Nicola Maaser

AARHUS BSS
DEPARTMENT OF ECONOMICS AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

Cooperation and Norm-enforcement under Impartial vs. Competitive Sanctions

Jan Philipp Krügel
Dept. of Economics & FOR 2104, Helmut-Schmidt University Hamburg

Nicola Maaser
Dept. of Economics and Business Economics, Aarhus University

October 26, 2020

Abstract

The willingness of mere bystanders, or "third parties," to incur costs to sanction non-cooperators in social dilemma situations has been documented in numerous studies. It is, however, not clear yet how different forms of higher-order punishment affect third party behavior and the level of cooperation. This paper experimentally studies incentives for third parties to enforce contribution norms in public-good games. We compare two treatments where the third party is embedded in different stylized institutions to a baseline treatment where this is not the case. In our treatments, the third party is, first, evaluated by another uninvolved individual ("fourth party"), and second, faces competition by another potential third party punisher. We find that third parties punish free-riding public good players more severely if they have to fear negative payoff consequences for themselves. Importantly, our results point to substantial qualitative differences between the institutional arrangements: When the third party is under scrutiny of a fourth party, punishment is more balanced, but also high compared to the other treatments. By contrast, competition between two third party candidates leads to strategic and partial punishment, generating the most profitable outcomes for public good players.

Keywords: third party punishment, higher-order punishment, cooperation, public goods game, experiments

JEL codes: C92; D02; H41

We have benefited from comments and suggestions by Stefan Traub and Fabian Paetzel, as well as participants at the GfeW meeting in Kassel and the ESA meetings in San Diego and Berlin. We gratefully acknowledge financial support by the University of Bremen's Central Research Development Fund.

1 Introduction

Sanctions imposed by objective outsiders, or third parties, play a crucial role in promoting and enforcing cooperative and norm-following behavior in economic and social interactions. The willingness to punish non-cooperators and at a personal cost is sometimes even considered the essence of social norms in contrast to instable retaliatory sanctions by individuals who are directly harmed by a norm violation (Kurzban et al. 2007). Numerous studies have documented this willingness experimentally (e.g., Fehr and Fischbacher 2004; Kamei 2017), through surveys (Traxler and Winter 2012), and in observational data (Mathew and Boyd 2011). Yet, whether third parties actually take action and to what extent will often depend on various factors such as such as their social proximity to the victim of a norm violation (e.g., Goette et al. 2006).

Societies have developed additional layers of rewards and punishments to incentivize third parties to mete out sanctions enforcing social norms. For example, a person who fails to join third-party sanctions against a norm violator can get ostracized himself by the community (see, e.g., Elster 1989, p.127). In firms, managers can implement sanctions for transgressive employees, but are themselves held to account for their decisions by their superiors or the supervisory board. In law enforcement, sentences by citizen-jurors have been found to be harsher after media reports of crime and shorter after reports on judicial errors (Philippe and Ouss 2018). One possible explanation is that media coverage makes crime more salient to jurors and raises their awareness of public sentiment towards these issues. An alternative institutional approach to achieve "good" third-party enforcement is to choose a third party by popular election. This is the case with judges and other judicial officers in some countries, but also commonly applies to religious communities and associations where members vote on their preferred candidate for leader. How do the incentives of third parties differ under these two frameworks, higher-level scrutiny and competitive election? Which is more effective in maintaining cooperation among unrelated individuals?

The present study addresses these questions experimentally. Different mechanisms to motivate and control a third party punisher have, to the best of our knowledge, not been investigated so far. By doing so, our study complements literature that examines third parties' underlying motivations in the absence of additional, institutional layers.

Our experimental design builds on a public goods game (PGG), a standard model of social dilemmas often faced by communities and work teams. Importantly, the existence of a norm for contribution behavior in the PGG is well established (see Bicchieri 2006). Contributions are observed by a third party who can choose to punish individual public goods players at a cost to himself. In addition to this baseline game, we vary the third party's incentives as follows: (i) We introduce another uninvolved outside player, referred to as "fourth party," who evaluates the decisions of the third party. (ii) We have two third parties who each submit a proposal for punishment; the proposal that receives the simple majority of the public good players' votes is implemented. In both these treatments, there are identical negative payoff consequences for the third party when others – the fourth

party or the majority of public good players – disapprove of their punishment proposal.

In line with previous evidence, we find that third parties are often willing to punish free riding behavior, even though punishment is costly for the third party and her monetary payoffs do not depend on the outcomes of the PGG. In the baseline game, where the third party is neither evaluated herself nor elected, her decisions can be interpreted as reflecting her intrinsic motivation to punish norm violators. The punishment can be rooted in inequity aversion (see Fehr and Schmidt 1999), or the desire to express negative feelings such as anger towards violators (see, e.g., Xiao and Houser 2005). Various evolutionary explanations have been suggested why such a propensity for altruistic punishment has developed (e.g., Boyd et al. 2003).[1]

We find that treatment variation has significant effects on punishment behavior. Third parties punish deviations from the contribution norm more severely if failing to do so or misjudging entails potential negative payoff consequences. The administered punishment appears most "balanced" when the third party is under observation of a fourth party. At the same time, the level of punishment is high compared to the other treatments. Electoral competition leads third parties to concentrate their punishment strategically on at most half of the public good players. This results in a relatively low average punishment and greater profits for public good players relative to the two other treatments.

Related literature — Perhaps closest to our approach is Kamei (2017) who compared the strength and effectiveness of third-party punishment between different decision-making formats. Specifically, he augmented the design frame of a prisoner's dilemma by including either an individual or a pair as a third party. An interesting finding is that an individual third party whose punitive actions are made known to another individual third party (attached to a different prisoner's dilemma dyad) is more effective than individual third parties left alone or pairs of third parties who decide jointly.

In a similar vein, Kurzban et al. (2007) examined how audience effects (see Filiz-Ozbay and Ozbay 2014) affect the sanctioning behavior of a third party towards players who exploited their partner in a trust game or in a sequential prisoner's dilemma. Specifically, administered punishments were low when the sanction remained completely anonymous compared to a condition where the experimenter got to know how much the third party had punished. Carpenter and Matthews (2012) studied a PGG where individuals in one group of players could observe and punish players of *another* group either in a one-directional or two-directional way. They concluded that norm enforcement by third parties can be driven by both indignation and group reciprocity. In a setting similar to our baseline, Engel and Zhurakhovska (2017) showed that framing plays an important role for third parties' behavior. Referring to them as "judges" or "public officials" instead of a neutral

[1]Note, however, that methodological concerns have been raised recently about the robustness of third-party punishment (Pedersen et al. 2018). These relate in particular to (i) experimenter demand effects and (ii) to establishing common knowledge about punishment possibilities between the third party and the potential recipients of punishment. Due to these issues it is not clear to what extent third party punishment in laboratory experiments should be interpreted as altruistic punishment. Yet, these issues do not pose major problems in our experiment where the focus is on how extra motivational layers affect third party punishment.

name tended to increase the balancedness of punishment decisions. A key difference with the above articles is that in our experiment the third party is not only being watched, but her decisions are subject to explicit scrutiny involving potential payoff consequences. This focus is relevant in light of incentive structures in some naturally-occuring settings, such as management and competition for office.

Second, our work is connected to studies comparing decentralized to centralized sanctions among directly affected individuals (second parties) with respect to cooperation and efficiency in social dilemma situations. A key finding is that schemes to punish non-cooperators are more effective when chosen democratically, by a majority of group members, than when they are exogenously imposed (Tyran and Feld 2006; Ertan et al. 2009; Dal Bò et al. 2010; Putterman et al. 2011; Markussen et al. 2014). Elections usually hand authority to individuals who refrain from antisocial punishment, i.e., who do not punish cooperative group members (Gross et al. 2016; see Herrmann et al. 2008 on the phenomeneon of antisocial punishment).

One line of research within this literature focuses on players' preferences for, and the endogenous formation of, sanctioning institutions. Nicklisch et al. (2016) let the players of a PGG choose beween a no-punishment environment, a centralized environment, where one randomly drawn subject is exclusively given the ability to punish, and a decentralized environment, where all players can punish. While centralized sanctions did not outperform decentralized sanctions in their experiment, they were nevertheless popular, especially when other players' contributions were not perfectly observable. Fehr and Williams (2018) suggested that centralized sanctions, by an elected judge, dominate decentralized environments when subjects have no opportunity to form a consensus about the contribution norm beforehand. While the above contributions consider various forms of second-party punishment, Kamei (2018) studied the strength of third-party sanctions in the context of a prisoner's dilemma when several third parties are present. He found that allowing third parties to determine the level of punishment democratically avoids both antisocial punishment and overly harsh punishment of norm violators.

Third, we contribute to the literature on higher-order punishment. Yet, the focus of most existing studies is on second-party punishment (in which the victim of a transgression can herself choose to punish the transgressor) rather than on third-party punishment. A key question then is whether non-punishers are punished by their peers (Cinyabuguma et al. 2006; Fu et al. 2017). Martin et al. (2019) compared the frequency of higher-order punishment against second and third parties, respectively, who *fail* to engage in punishment. They found that higher order sanctions are more common against non-punishing third parties than against non-punishing second parties, which points to the injunctively normative nature of third party punishment.

The remainder of this paper is organized as follows. The next section presents our theoretical framework. Section 3 describes our experimental design. The experimental results are reported in Section 4. We conclude in Section 5 and provide additional materials in two appendices.

2 Theoretical framework

In this section, we sketch a simple framework to help motivate our predictions about norm enforcement in a PGG. The baseline game consists of two stages. Stage I is a standard PGG in a group of n players. The contributions in the PGG are observed by a third-party player who, in stage II, has the option to punish the players of the PGG.

In stage I, each public good player i has an endowment y and chooses, simultaneously with other subjects, how much to contribute to a public good with marginal per capita return α with $0 < \alpha < 1 < n\alpha$. The public good player may be awarded p_i punishment points which reduce his payoff by βp_i ($\beta > 0$). Monetary payoff thus is

$$\pi_i = y - c_i + \alpha \sum_{j=1}^{n} c_j - \beta p_i. \tag{1}$$

The restriction $0 < \alpha < 1$ implies that $\partial \pi_i / \partial c_i < 0$, i.e., contributions to the public good reduce individual monetary payoff all else equal. However, the restriction $1 < n\alpha$ implies that $\partial \sum_i \pi_i / \partial c_i > 0$ so that aggregate group payoff would be maximized if each individual contributed the whole endowment to the public good.

Following Krupka and Weber (2013) and Gächter et al. (2017), we assume that individuals are motivated by both material self-interest and a preference for adhering to norms, i.e., collectively recognized rules of behavior that define which actions are viewed as appropriate. $N(\mathbf{c}) : [0, y]^N \to [-1, +1]^N$ is a norm that indicates for each level of contribution whether it was appropriate ($N(\cdot) > 0$) or inappropriate ($N(\cdot) < 0$). The norm can depend on context such as the the level of anonymity in the group or be a norm of conditional cooperation, which require cooperation if others cooperate, but allows defection if others defect.[2] We assume that individuals' preference for norm conformity entails the wish to see norm violations punished. Individual i's utility function is given by

$$U_i = \pi_i + \gamma_i V(N(\mathbf{c}), \mathbf{p}) \tag{2}$$

where $\mathbf{p} = (p_1, \ldots, p_n)$ is the punishment profile. Function $V(\cdot)$ represents the common assesment of how appropriate contributions and punishments are. The parameter $\gamma_i \in [0, 1]$ measures the extent to which individual i has internalized the norm; if γ_i is greater, i derives greater utility from norm-following behavior herself as well as attaching greater importance to others' norm compliance and norm enforcement. We treat γ_i as an innate individual trait. A player i with $\gamma_i = 0$ does not care about norms; she might nevertheless choose a non-zero contribution level if this is necessary to maximize payoffs in face of potential punishment. We assume that $V(\cdot)$ admits a unique maximum at the point where behavior is fully consistent with the relevant norm and no punishment is administered. Otherwise, we assume that

$$\left.\frac{\partial V(N(\mathbf{c}), \mathbf{p})}{\partial p_i}\right|_{N(c_i)<0} > 0 \qquad \text{and} \qquad \left.\frac{\partial V(N(\mathbf{c}), \mathbf{p})}{\partial p_i}\right|_{N(c_i)\geq 0} < 0, \tag{3}$$

[2] Elster XXXX labels this a quasi-moral norm.

i.e., punishment of player i increases $V(\cdot)$ if i's contribution level was inappropriate, and decreases $V(\cdot)$ if i's contribution was appropriate under norm $N(\cdot)$. Moreover, we assume that

$$\left.\frac{\partial^2 V(N(\mathbf{c}),\mathbf{p})}{\partial p_i^2}\right|_{N(c_i)<0} < 0. \tag{4}$$

Similar to public good players, the third party (TP) has monetary utility $\pi_{TP} = y_{TP} - \sum_{j=1}^{n} p_j$ and a preference $\gamma_{TP} \geq 0$ for norm compliance. Her utility is given by

$$U_{TP} = \pi_{TP} + \gamma_{TP} V(N(\mathbf{c}),\mathbf{p}). \tag{5}$$

The third party chooses the vector of punishments that maximizes (5). Clearly, if γ_{TP} equals zero, TP will not punish. Otherwise, she will optimally award punishments to public good players whose contributions were inappropriate. The first-order condition is

$$\frac{\partial U_{TP}}{\partial p_i} = -1 + \gamma_{TP}\frac{\partial V(N(\mathbf{c}),\mathbf{p})}{\partial p_i} = 0. \tag{6}$$

Assuming that an interior solution exists, TP's optimal punishment to player i is implicitly defined by

$$p_i^{\star} = V_{p_i}^{-1}\left(N(\mathbf{c}), \frac{1}{\gamma_{TP}}\right). \tag{7}$$

By our assumptions (3) and (4) above, it follows from (7) that TP will punish more, the higher γ_{TP} is. Generally, the TP's punishment decisions weigh the cost of using (part of) her endowment for punishing a public good player against the cost of not following her preferences for norm compliance.

Now suppose that a fourth party (FP) has the power to reduce TP's payoff at no cost to himself after observing outcomes of the PGG and TP's punishment profile. TP's utility then becomes

$$U_{TP|FP} = \begin{cases} \pi_{TP} + \gamma_{TP} V(N(\mathbf{c}),\mathbf{p}) - \delta & \text{if FP chooses to punish TP} \\ \pi_{TP} + \gamma_{TP} V(N(\mathbf{c}),\mathbf{p}) & \text{if FP does not punish TP} \end{cases} \tag{8}$$

where δ is the reduction in TP's payoff if FP disagrees with the punishment profile. As punishment is costless to her, a FP (with any $\gamma_{FP} > 0$) wants to see a punishment profile such that

$$\frac{\partial V(N(\mathbf{c}),\mathbf{p})}{\partial p_i} = 0 \quad \forall i.$$

When is TP willing to change her optimal punishments $\mathbf{p}^{\star}$ defined by (7) to punishment profile $\mathbf{p}^{\star\star} > \mathbf{p}^{\star}$? This will be the case if

$$y_{TP} - \sum_{j=1}^{n} p_j^{\star\star} + \gamma_{TP} V(N(\mathbf{c}),\mathbf{p}^{\star\star}) \geq y_{TP} - \sum_{j=1}^{n} p_j^{\star} + \gamma_{TP} V(N(\mathbf{c}),\mathbf{p}^{\star}) - \delta$$

$$\Leftrightarrow \qquad \delta \geq \gamma_{TP}[V(N(\mathbf{c}),\mathbf{p}^{\star}) - V(N(\mathbf{c}),\mathbf{p}^{\star\star})] + \sum_{j=1}^{n} p_j^{\star\star} - \sum_{j=1}^{n} p_j^{\star}.$$

The right-hand side of this last inequality is always positive due to concavity of $V(N(\mathbf{c}), \mathbf{p})$ in $\mathbf{p}$. Thus, the presence of FP causes TP to change her punishment profile only if δ is 'large enough'.

Finally, consider the possibility that public good players select a TP by voting on two candidates. Each candidate submits a punishment proposal after observing the PGG. The candidate who receives a majority of votes is elected and her punishment proposal is implemented. Candidate k's utility thus is

$$U_k = \begin{cases} \pi_{TP} + \gamma_k V(N(\mathbf{c}), \mathbf{p}_k) & \text{if } k \text{ is elected} \\ 0 & \text{if } k \text{ is not elected.} \end{cases} \tag{9}$$

A public good player will vote for candidate k if her utility is greater under $\mathbf{p}_k$ than under candidate l's proposal $\mathbf{p}_l$; she will vote for either candidate with equal probability if she is indifferent between the two proposals. She thus prefers, other things equal, the punishment profile that matches his γ_i more closely and punishes players who made lower contributions than herself. The decisive voter is the individual whose preference γ^M for norm conformity is the median in her group of public good players. The candidate whose proposal corresponds most closely to these preferences is elected. TP candidates infer the median public good player's parameter γ^M from her contribution level. Let $\mathbf{p}^{\star\star\star}$ denote the optimal punishment profile of a player i with $\gamma_i = \gamma^M$. Then TP candidates will both propose $\mathbf{p}^{\star\star\star}$, which involves punishment for at most $(n-1)/2$ public good players whose contributions fell short of the median individual's contribution.

3 Experimental design and hypotheses

3.1 The public good game

The Baseline treatment is a standard PGG which is repeated for 20 periods. At the beginning of the first period, all participants are randomly assigned to one of two roles – public good player ("A-Player") or third party ("B-Player").[3] Roles stay fixed for the entire 20 periods. In each period three A-Players (A_1, A_2 and A_3) are randomly matched with one B-Player using a stranger matching protocol. We repeat the game for 20 periods to allow for learning and changes of behavior over time. The random matching procedure avoids individual reputation building.

At the beginning of every period, each A-Player receives an endowment of 20 points and B-Players get 30 points. The A-Players are then asked to decide how many points $c_i \in \{0, 4, 8, 12, 16, 20\}$ they want to contribute to a public good with a marginal per capita return $\alpha = 0.5$, which is implemented in their group. All contributions are made simultaneously and without communication. After the PGG players have made their decisions, the third party is informed about the contributions of A_1, A_2 and A_3 in her group. The

[3]Note that we use a neutral framing ("A-Player", "B-Player" and "public project") in the experimental instructions.

third party can then punish these A-Players by assigning punishment points p_i. One punishment point reduces the payoff of A-Players by two points, but costs the B-Player only one point herself. The monetary payoff π_{A_i} for A-Player i is given by

$$\pi_{\mathrm{A}_i} = 20 - c_i + 0.5 \sum_{j=1}^{3} c_j - 2p_i. \tag{10}$$

A-Players cannot receive a negative payoff. If the formula yields a negative amount, the payoff is 0 points. The monetary payoff for a B-Player in a BASELINE period equals her endowment of 30 points minus the punishment points she assigned to A_1, A_2, and A_m:

$$\pi_{\mathrm{B}}^{Base} = 30 - \sum_{i=1}^{3} p_i. \tag{11}$$

At the end of each period, the participants receive feedback about all decisions of their group members and they are also informed about their own payoff in the period. The final payoff for each participant in the PGG is calculated as the sum of her payoffs over the 20 periods.

In the second treatment FOURTH PARTY, we introduce a "C-Player" into the framework. Each group in FOURTH PARTY consists of three A-Players (A_1, A_2 and A_3), one B-Player and one C-Player. Roles stay fixed across 20 periods and groups are randomly rematched in every round. The first two stages are the same as in BASELINE: First, the public good players decide on their preferred contribution to the public good; then, the third party is informed about the A-Players' decision and can assign punishment points. Afterwards, in a third stage, the C-Player receives information on the choices of the B-Player in her group. The C-Player is then asked if he "agrees" or "disagrees" with the punishment of the B-Player. Each C-Player receives a fixed payoff of 15 Euros irrespective of his answer.[4] The payoffs for A-Players are calculated by (10), i.e., in the same way as in BASELINE. The payoff for the B-Player, however, now depends on C's decision:

$$\pi_{\mathrm{B}}^{Fourth} = \begin{cases} 30 - \sum_{i=1}^{3} p_i & \text{if C agrees} \\ 5 & \text{if C disagrees.} \end{cases} \tag{12}$$

In the third treatment COMPETITION, each group consists of three A-Players as before and two B-Players (B_1 and B_2). Again, roles stay fixed across 20 periods and groups were randomly rematched in every period. In the first stage, the three public good players decide on how much to contribute to the public good. In the second stage, the two B-Players receive information about the decision of the A-Players in their group and both B-Players can assign punishment points to A_1, A_2 and A_3. In the third stage, the A-Players are informed about the punishment proposals of B_1 and B_2. They then indicate their preferred proposal by vote, and the proposal which receives the majority of votes is implemented.

[4]The payoff for C is designed in a way that it approximately equals the average final payoff of A- and B-Players.

The payoff for an A-Player is calculated by (10), using the winning punishment proposal. A B-Player's payoff now depends on the decision of the three A-Players in her group. The selected B-Player's payoff is calculated as in (11); the non-selected B-Player receives a payoff of 5 points:

$$\pi_{\mathrm{B}}^{Pol} = \begin{cases} 30 - \sum_{i=1}^{3} p_i & \text{if B's proposal received two or three votes} \\ 5 & \text{if B's proposal received zero or one vote.} \end{cases} \tag{13}$$

3.2 Procedures

The experiment was programmed using z-Tree (Fischbacher 2007). The participants were recruited via the administration software hroot (Bock et al., 2014). As participants might differ with respect to their inequality aversion and efficiency preferences, and this might in turn influence their punishment and contribution decisions, we also conducted the equality equivalence test due to Kerschbamer (2015) to elicit these preferences in a separate part of the experiment. In order to save space, we omit details and refer to the original description of the double price-list technique in that paper.[5] Subjects completed the Kerschbamer-test first (part 1), before playing the PGG (part 2).[6] As final payoff, each participant received the sum her individual payoffs from parts 1 and 2 at a conversion rate of 100 points = 3 Euros. The subjects answered some control questions after reading the instructions and completed a questionnaire upon conclusion of the experiment.

The experiment was conducted at the experimental laboratory of the University of Hamburg and involved six sessions with a total of 168 participants. We had two sessions in each treatment: 48 subjects participated in Baseline (24 subjects per session) and 60 subjects (30 subjects per session) participated, respectively, in Fourth Party and Competition. Upon arrival at the laboratory, the participants were randomly placed at the computers. For each of the two parts of the experiment they received written instructions, which were read aloud by the experimenter. Sessions lasted for about 75-90 minutes. The highest payoff was €20.52, the lowest payoff €6.93 and the average payoff €15.48. All decisions and payoffs were made in private.

3.3 Hypotheses

We can derive the following hypotheses from the theoretical framework presented in Section 2:

Hypothesis 1 (How much punishment?). *Punishment per unit of norm violation in the* Fourth Party *treatment will (weakly) exceed that in the* Baseline *treatment.*

[5]The test provided two measures of distributional preferences: (i) a measure of inequality aversion, the willingness-to-pay for advantageous inequality $WTP^a \in [-0.667, 0.667]$; (ii) a measure for efficiency preferences, the willingness-to-pay of disadvantageous inequality, $WTP^d \in [-0.667, 0.667]$.

[6]The order of the experimental parts was chosen this way so that participants could familiarize themselves with the experimental environment before completing the main task.

Hypothesis 2 (Who gets punished?).

(a) In the Competition *treatment, both candidates propose to punish exactly one public good player, the least contributor. In case that all contributions are identical, no punishment is allocated.*

(b) The number of public good players who get punished exhibits greater variation in the Baseline *and the* Fourth Party *treatments compared to the* Competition *treatment.*

Hypothesis 3 (Success of third parties).

(a) In the presence of a Fourth Party*, the probability that third party decisions meet with approval increases in the assigned punishment.*

(b) If a candidate in the Competition *treatment proposes to punish more than one public good player, his proposal does not win against a competitor who allocates punishment in line with Hypothesis 2(a).*

The theoretical framework is static and has nothing to say about how public good players adapt to received punishment. But also in a static setting, we still expect players to understand third parties' incentives in the three conditions and take that into account when maximizing (2):

Hypothesis 4 (Amount of contributions). *Public good players anticipate to be punished more if a* Fourth Party *is present compared to the* Baseline *treatment; the temptation to free ride is thus diminished.*

4 Results

We will first provide a brief overview of our main results before continuing with a more detailed analysis of punishment behavior, contributions and earnings.

4.1 Overview

To first get a sense of contribution behavior, Figure 1 shows average contributions to the public good by treatment (left panel) and by treatment and period (right panel). Moreover, we report significance levels of bilateral treatment comparisons based on Wilcoxon ranksum tests in Table 1. In total, we have 720 observations for contribution decisions in each treatment.[7] A first finding is that both Fourth Party (mean contribution: 10.64 points) and Competition (mean contribution: 9.44 points) generate significantly higher contribution rates than Baseline (mean contribution: 6.59 points). In the Baseline

[7]We pool the data for each individual as an independent observation for the Wilcoxon test reported in Table 1, since the game is played for 20 periods.

treatment, we also see a slight, negative contribution trend over periods, which cannot be observed as clearly in Fourth Party and Competition. The data replicate the stylized fact from past PGG experiments that on average, participants initially contribute between 40 % and 60 % of their endowment (see Chaudhuri 2011).

Fig. 1. Mean contribution by treatment and over time

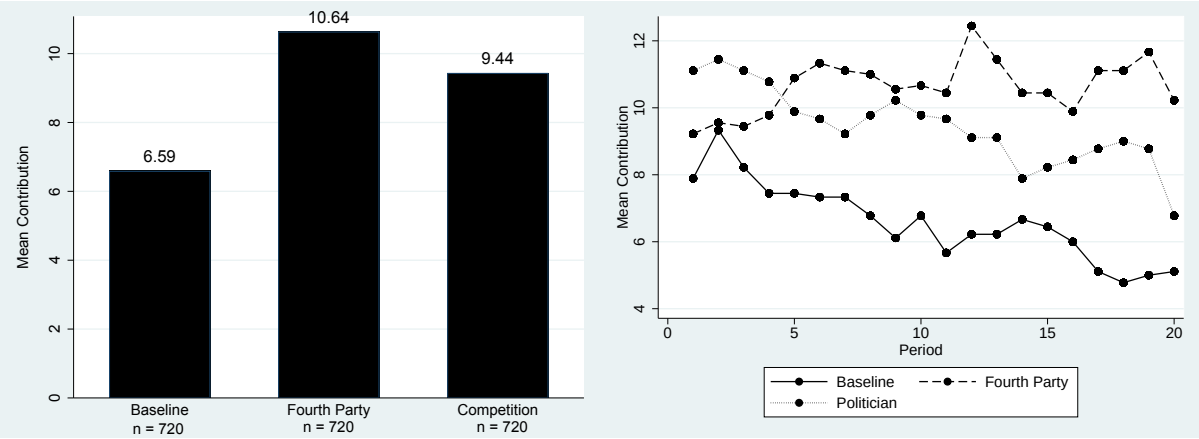


Table 1.
Bilateral treatment comparisons

			p-values of bilateral comparison	
Variable	Treatment	Mean	*Baseline*	*Fourth Party*
	Baseline	6.59		
Contribution	*Fourth Party*	10.64	.001	
	Competition	9.44	.003	.421
	Baseline	1.00		
Punishment	*Fourth Party*	1.65	.028	
	Competition	1.03	.356	.020

Notes: Mean of contributions and punishment across the 20 periods and p-values of Wilcoxon ranksum tests for bilateral treatment comparisons. All tests based on means of individual contribution/punishment averages (n = 36 for contributions in all treatments; n = 12 for punishment in Baseline and Fourth Party and n = 24 for punishment in Competition.)

In the second stage of each treatment, third parties were able to punish PGG players. Figure 2 shows average punishment points assigned by treatment (left panel) and by treatment and period (right panel).[8] Average punishment was significantly higher in Fourth Party (mean punishment: 1.65 points) than in Competition (mean punishment: 1.03

[8]In the treatment Competition, there were two third parties in each group. Although only one punishment proposal was implemented, we incorporate both punishment proposals when calculating the mean in Figure 2 and Table 1.

points) and BASELINE (mean punishment: 1.00 points) (see also lower part of Table 1). We also observe that in 1895 of a total of 2880 individual punishment decisions (65.8%) in all treatments, third parties chose not to punish. Many public good experiments with punishment show decreasing punishment levels over time. Interestingly, we do not see a clear punishment trend in any of the treatments of our experiment (see Figure 2, right panel).

Fig. 2. Mean punishment by treatment and over time

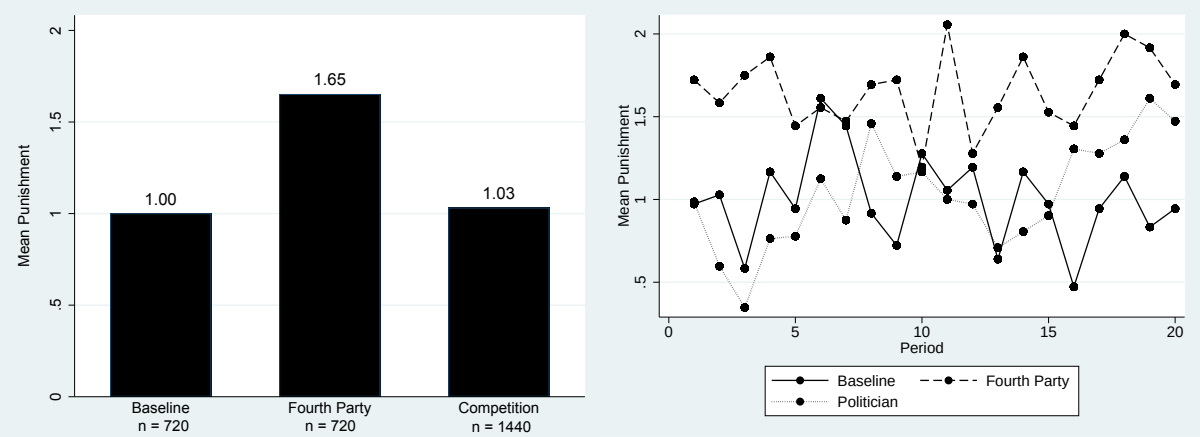
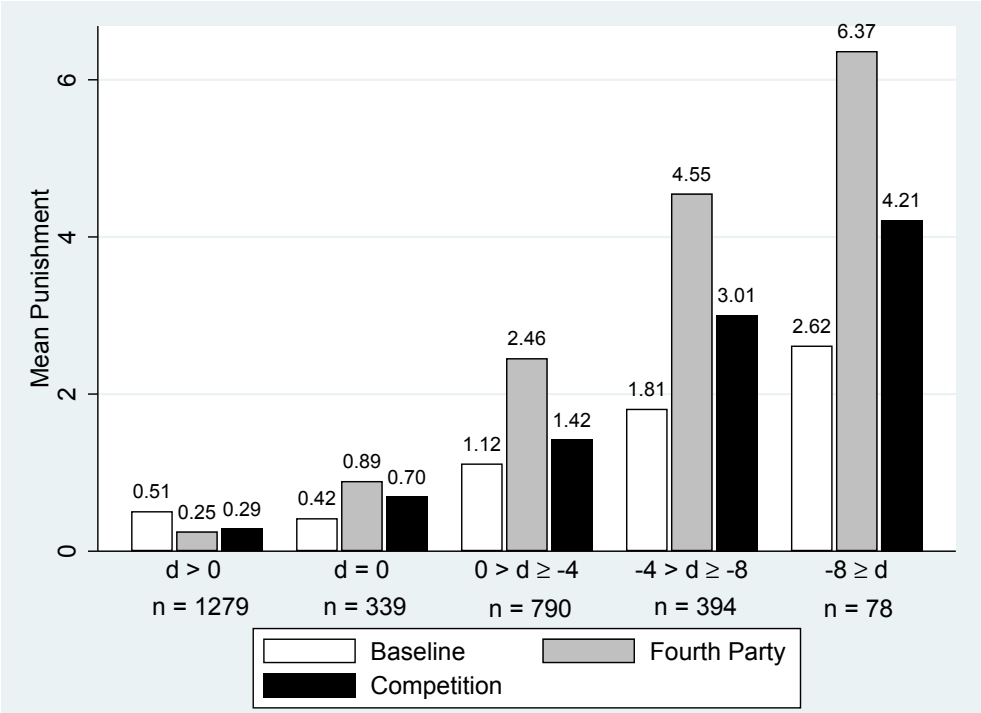


Figure 3 shows how PGG players' contributions affected the third parties' punishment decisions. The graph reports the mean punishment of third parties by deviation (d) of a public good player's contribution from the average contribution in his group. We categorize the deviation by intensity level and by treatment. The bars in the two lowest d-categories indicate that a few public good players were punished even though they contributed exactly at the group average ($d = 0$) or more than group average ($d > 0$). However, mean punishment in these cases was generally at a low level in all three treatments. Punishment clearly increased when the contribution of public good players negatively deviated from the group average. We also oberve a stronger effect for greater deviations. The three groups of bars on the right of Figure 3 indicate that third parties punished negative deviations most severely in the FOURTH PARTY treatment.

4.2 Third party behavior

Our main interest is to analyze the effects of variation in third parties' incentives. We run several regressions to identify the key factors that determine the amount of assigned punishment. Since we have 1895 observations of third parties with no punishment, we analyze the likelihood of punishment or not separately from the severity of punishment. Table 2 presents random effects probit regressions in columns (1) and (2) where the dependent variable takes a value of 1 if the B-Player punished an A-Player and 0 otherwise. Columns (3) and (4) include random effects regressions where the dependent variable is the

Fig. 3. Deviation from average group contribution and individual punishment



Notes: d = individual deviation from the average contribution in the group

level of third party punishment (truncated at zero). We use Baseline as our benchmark treatment in all regressions. Standard errors are clustered at the individual level.

The first regression suggests that public good players chose to punish more often in Fourth Party than in Baseline and Competition. Another significant effect on the punishment decision relates to the deviation of an individual's contribution from the average contribution in his group. We observe that a positive deviation from average group contribution leads to significantly less punishment. In regression (2), we include two interaction terms: *Fourth Party* x *deviation* and *Competition* x *deviation*. Both terms have a significant and negative influence on punishment. Thus a treatment difference in assigned punishment only seems to occur if the contribution of a public good player (negatively) deviates from group average, an observation that is consistent with the results presented in Figure 3. The interaction terms show that the strongest effect for *deviation* on the punishment decision was present in the Fourth Party treatment. A χ^2-test indicates that the difference between the interaction terms *Fourth Party* x *deviation* and *Competition* x *deviation* is statistically significant ($p = 0.058$; compare last row of Table 2): The punishment decision was more responsive to deviation from the average in Fourth Party and those who contributed less than their group members were consistently punished by third parties. In regression (2), we include the standard deviation of group contributions and period indicators as additional controls. We also test for the inequality aversion (WTP^a)

and efficiency preferences (WTP^d) of third parties, which we elicited separately with Kerschbamer's test in the first part of the experiment. However, these control variables are all insignificant.

In regressions (3) and (4), we analyze the punishment level. The basic treatment effects are insignificant in both regressions. Regression (4) shows that negative deviations of public good players from average group contribution are punished most severely in Fourth Party. A χ^2-test confirms that the difference between the interaction terms *Fourth Party* x *deviation* and *Competition* x *deviation* is significant ($p = 0.057$). The regression additionally reveals a significant positive impact of more spread out contributions on the punishment level. In contrast to regression (2), period effects are significant in (4). Generally, the indicators for the punishment level are similar to those for the punishment decision. We therefore conclude:

Result 1. *If the contribution of public good players negatively deviates from group average, third parties punish more consistently and more severely in* Fourth Party *than in* Baseline *(and in* Competition*), thus confirming Hypothesis 1.*

We now turn to third party punishment at the group level. Specifically, we analyze *how many* public good players were punished in the different treatments.[9] Looking at all treatments combined, the majority of third parties tended to punish either none or only one of the public good players. In case only one player was punished, the punishment was almost always directed at the player who contributed least to the public good – this was true in 98.32% of cases. We see similar results when two players were punished. In this case, punitive action was directed at the lowest two contributors to the public good in 97.97% of cases.

To get a more comprehensive picture, Figure 4 displays how many of the three public good players in a group were punished (in percent, by treatment and pooled over all 20 periods). In accordance with Figure 2, we observe that in Baseline, third parties punished considerably less than in the other treatments. In fact, 51.67% of the time none of the A-Players in a group were punished. The picture is very different in Fourth Party, where punishment was not only more severe (cf. Figure 2) but was also often directed at several public good players: Only 16.25% of time no player was punished. Punishment for one or two players was much more common with 33.33% and 43.75% of cases, respectively. With respect to the Competition treatment, Figure 4 also displays punishment profiles for 'rejected' and 'accepted' third parties, i.e., whose proposals respectively received the minority and majority of public good players' votes. We separate the two graphs in order to show the influence of the punishment profile on the electoral success of third parties in this treatment.[10] In line with our Hypothesis 3(b), third parties seem to be more successful in winning the majority of votes when punishing at most one of the three public good players.

[9] In line with our estimates reported in Table 2, we also conducted regressions at the group level for the punishment decision and the punishment level (reported in Table B1, Appendix B). In these regressions, we also included lagged variables as further controls.

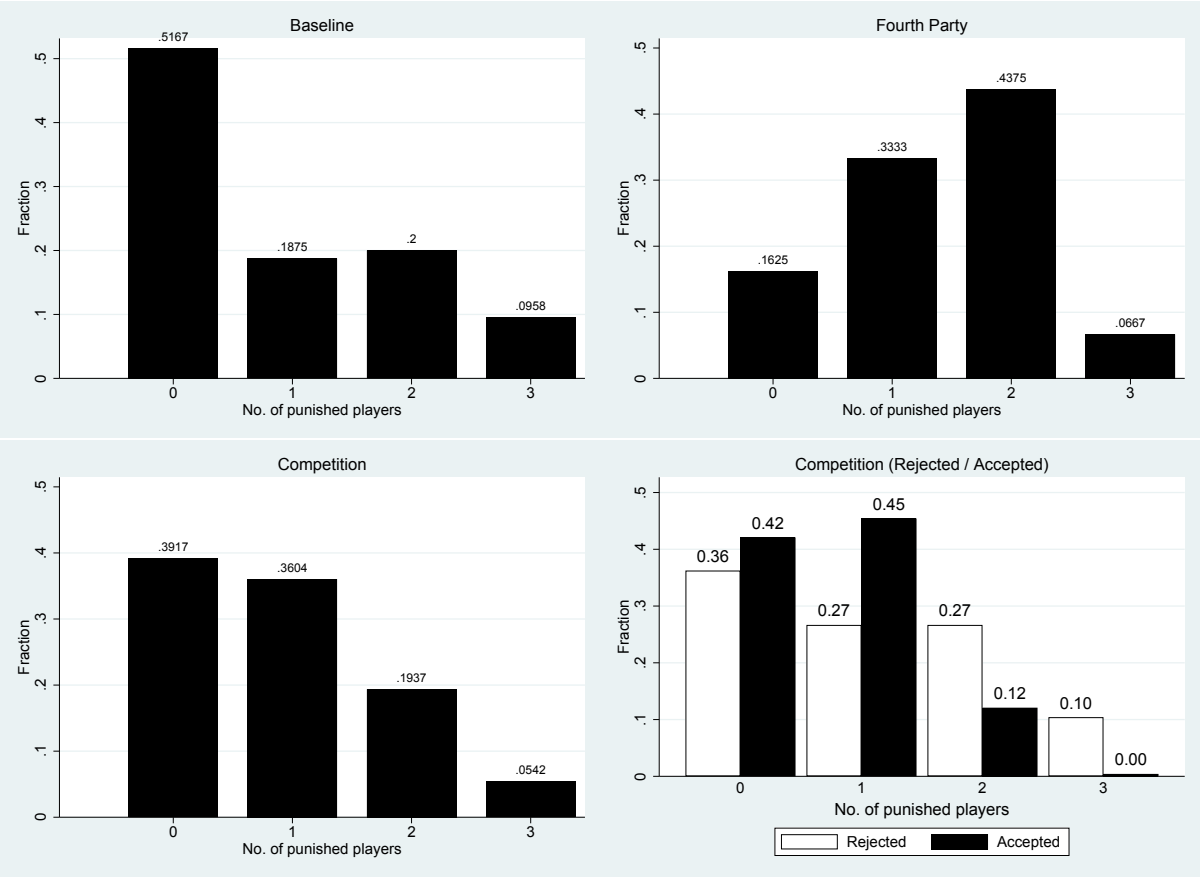
[10] A figure that separates punishment profiles for third parties by assessment of the fourth party in the treatment Fourth Party can be found in Appendix B.

Table 2.
Estimates for individually assigned punishment

Variable	(1) Punishment decision	(2) Punishment decision	(3) Punishment level	(4) Punishment level
Fourth Party	1.137*	0.854	0.886	-0.322
	(0.641)	(0.670)	(0.762)	(0.719)
Competition	0.487	0.223	0.752	0.202
	(0.568)	(0.555)	(0.746)	(0.781)
Deviation	-0.244***	-0.149***	-0.325***	-0.134***
	(0.028)	(0.032)	(0.058)	(0.049)
Fourth Party x Deviation		-0.245***		-0.321***
		(0.079)		(0.091)
Competition x Deviation		-0.089*		-0.154**
		(0.051)		(0.068)
SD of group contributions		0.040		0.112***
		(0.027)		(0.044)
Period		0.003		0.049**
		(0.009)		(0.022)
WTP^a		0.759		0.833
		(0.789)		(1.183)
WTP^d		0.200		-0.092
		(0.506)		(0.904)
Constant	-1.341**	-1.540**	1.401*	0.758
	(0.560)	(0.637)	(0.753)	(0.822)
Observations	2880	2880	985	985
Wald-χ^2	78.97***	106.68***	50.87***	130.47***
Fourth Party x Deviation = *Competition* x Deviation		$p = 0.058$		$p = 0.057$

Notes: (1) and (2): Random-effects probit regressions where the dependent variable takes a value of 1 if the B-Player punished an A-Player and 0 otherwise. (3) and (4): Random-effects regressions where the dependent variable is the number of punishment points a B-Player assigned per A-Player; the dependent variable in (3) and (4) is truncated at zero. Standard errors, clustered at the individual level in all 4 regressions, in parentheses. Deviation: Individual contribution minus average contribution within the group. SD: Standard deviation. *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

Fig. 4. Number of punished players by treatment (fractions)



In Table 3, we report logit regressions where the dependent variable is an indicator variable that captures how many public good players were punished by each third party. In column (1) the indicator variable equals one if the number of punished A-players in the group was zero, and it equals zero if punishment occurred. We use the Baseline treatment again as a benchmark. The regression shows a negative and significant impact of the treatment variables *Fourth Party* and *Competition*, which implies that 'no punishment' was more frequent in Baseline than in the other treatments. Abstaining from punishment was also more common in Competition than in Fourth Party. Moreover, a high standard deviation in group contributions had a negative and significant impact. That is, when the level of contributions differed greatly within a group, 'no punishment' was rare.

In column (2) the indicator equals one when exactly one player was punished, and zero otherwise. Clearly, one player was punished more often in Competition than in Baseline. This result confirms our Hypothesis 2(b).[11] In contrast to the first regression, the coefficient for the standard deviation of contributions in 'one player punished' is positive.

[11]Note that the variable *Competition* in Table 3 does not separate between successful and unsuccessful third parties.

Table 3.
Number of punished players in a group: regressions

Variable	(1) no punishment	(2) one player punished	(3) >one player punished
Fourth Party	-3.180***	1.321**	1.475*
	(1.098)	(0.652)	(0.811)
Competition	-1.381	1.548**	0.015
	(0.905)	(0.632)	(0.817)
SD of group contributions	-0.273***	0.109**	0.114***
	(0.059)	(0.043)	(0.042)
Period	-0.034	0.022	-0.000
	(0.024)	(0.018)	(0.016)
Constant	2.391**	-2.971***	-2.140**
	(1.041)	(0.771)	(0.836)
Observations	960	960	960
Wald-χ^2	28.267***	9.094*	15.846***

Notes: Random-effects logit regressions. Dependent variable: 0 players punished ($yes = 1/no = 0$) / 1 player punished ($yes = 1/no = 0$) / >1 player punished ($yes = 1/no = 0$). Standard errors, clustered at the individual level, in parentheses. SD: Standard deviation. $^*p \leq 0.10$. $^{**}p \leq 0.05$. $^{***}p \leq 0.01$.

The third regression '>one player punished' shows that two or three players were punished significantly more often in Fourth Party than in Baseline and Competition. Overall, these finding reinforce the following result:

Result 2. *Punishment in* Fourth Party *is often directed at several players, whereas punishment in* Competition *is often strategically directed at only the biggest deviator from the contribution norm. This confirms Hypotheses 2(a) and 2(b).*

We next take a closer look at treatments Fourth Party and Competition in order to see whether these two central findings are supported by further analyses. In Fourth Party, there are 240 cases where a fourth party player had to evaluate the punishment decision of third parties. The assessment was often positive: In 80.42% of cases, the fourth party agreed with the punishment decision of the third party.

Column (1) in Table 4 presents a random effects regressions where a dummy ('FP agrees') indicating a positive assessment by the fourth party is the dependent variable. As explanatory variable, we include "Punishment of norm deviators", which equals one if a third party punished *all* public good players who contributed less than group average, and zero otherwise. The regression reveals that this variable does not have a significant effect on the assessment of fourth parties. A higher average punishment within the group, on the other hand, does have a positive and significant effect on receiving approval. The results indicate that in order to be positively assessed in Fourth Party, it is more important to punish extensively rather than to simply punish those who contribute less than group

average or less than half of their endowment.[12]

Column (2) in Table 4 reports a regression that has the electoral success of third parties in COMPETITION as the dependent variable. Just like in the first regression, the 'Punishment of norm deviator'-variable does not significantly influence electoral success in COMPETITION. However, the dummy for having one public good player punished has a significant positive effect on electoral success.[13] Furthermore, a higher average punishment within a group led to significantly reduced electoral success. We conclude:

Result 3. *Strategic punishment of only one deviator is key to third party candidate success in* COMPETITION*, whereas a high average punishment is harmful. In* FOURTH PARTY*, a high average punishment led to greater approval by fourth parties. This confirms Hypotheses 3(a) and 3(b).*

Table 4.
Success of third parties in FOURTH PARTY and COMPETITION

Variable	(1) FOURTH PARTY	(2) COMPETITION
Punishment of norm deviators	-0.085	-0.151
	(0.442)	(0.326)
Average punishment within group	0.703***	-0.418***
	(0.224)	(0.155)
One player punished	0.069	1.030***
	(0.251)	(0.264)
Constant	0.536	0.088
	(0.543)	(0.192)
Observations	240	480
Wald-χ^2	17.258***	33.649***

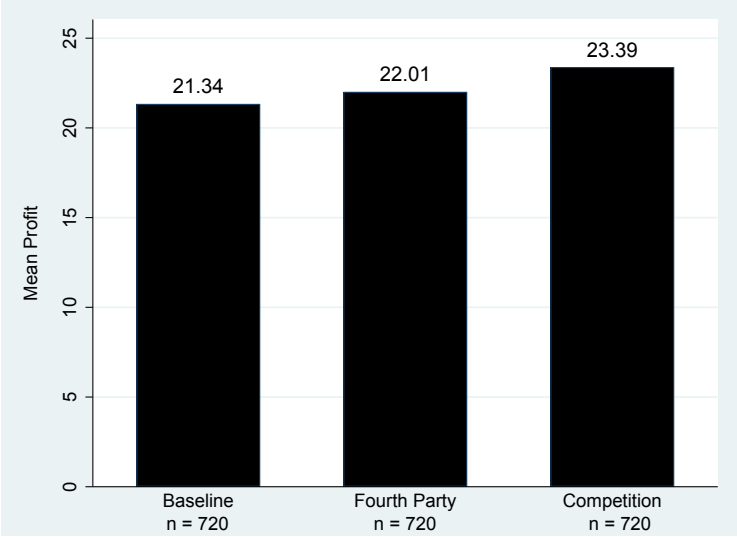
Notes: Random-effects logit regressions. Dependent variable: Positive assessment of third party by fourth party in (1) ($yes = 1/no = 0$) / Electoral success, i.e., third party received majority of votes from public good players in (2) ($yes = 1/no = 0$). Standard errors, clustered at the individual level, in parentheses. Punishment of norm deviators: third party punished all public good players who contributed less than group average ($yes = 1/no = 0$). *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

[12]Different players might have differing contribution norms in mind or have different preferences for norm compliance. With this in mind, we conducted additional regressions using the sum of the (squared) deviation of third parties' punishment to an 'ideal' punishment profile as another explanatory variable. However, the choice of the ideal punishment profile is to some extent arbitrary. We therefore relegate corresponding results to Appendix B (see Table B2 there).

[13]As we mentioned earlier, in case only one player was punished, the punishment was almost always directed at the lowest contributor to the public good. Results in columns (1) and (2) are virtually identical when exchanging 'One player punished' with a dummy that also captured whether this player was also the lowest contributor to the public good (see Appendix B, Table B2).

4.3 Contributions and earnings

Fig. 5. Earnings of public good players by treatment



Finally, we turn to the behavior of the public good players. It became already clear from Figure 1 that the level of contributions largely depended on the treatment. For a more detailed analysis, we conducted three random effects tobit regressions of total contributions (see columns (1)-(3) in Table 5). The regressions confirm that contributions were significantly higher in Fourth Party and Competition than in the benchmark Baseline treatment. χ^2-tests indicate that the difference between *Fourth Party* and *Competition* is not statistically significant (cf. last row of Table 5). In addition, regression (2) shows that 'period' has a significant negative impact on contributions. In regression (3), we also include two lagged variables. Interestingly, a higher average contribution of the other group members in the previous period has a significant positive effect on contributions, while the coefficient for received punishment in the previous period is insignificant.[14] These results suggest that there were conditional cooperators in our sample who (partly) based their contribution decision on observed behavior of others.

Lastly, we analyze which treatment generated the highest earnings for public good players. Figure 5 displays mean profits of public good players by treatment. We do not consider earnings of third parties in our analysis because these were influenced by our parametrization. Moreover, third parties are "outsiders" in our scenario in the sense that they do not benefit from high contributions to the public good. In contrast, the earnings of the public good players are comparable across treatments. Figure 5 shows that average earnings were highest in Competition, followed by Fourth Party and Baseline. Wilcoxon ranksum tests with pooled data show that the differences in mean

[14]We did not include received punishment of the current period because punishment decisions of third parties occurred *after* public good players had made their contributions.

Table 5.
Contributions and earnings of public good players

Variable	(1)	(2)	(3)	(4)
	Contributions			Earnings
Fourth Party	6.602***	6.616***	5.929***	0.672
	(2.053)	(2.049)	(2.067)	(0.594)
Competition	4.626**	4.644**	3.830*	2.050***
	(2.045)	(2.042)	(2.059)	(0.604)
Period		-0.133***	-0.125***	-0.054***
		(0.019)	(0.020)	(0.018)
Punishment received $(t-1)$			-0.016	
			(0.064)	
Average contributions of others $(t-1)$			0.230***	
			(0.030)	
Constant	4.511***	5.887***	4.308***	26.302***
	(1.463)	(1.474)	(1.504)	(1.097)
Observations	2160	2160	2052	2160
Wald-χ^2	10.866***	58.069***	123.978***	30.03***
Fourth Party = Competition	$p = 0.330$	$p = 0.330$	$p = 0.304$	$p = 0.000$

Notes: Random-effects tobit regressions in (1)-(3)/ Random-effects regression in (4). Dependent variable: Contributions in (1), (2) and (3)/ Earnings in (4). Standard errors in parentheses. In (4), standard errors are clustered at the individual level. $^*p \leq 0.10$. $^{**}p \leq 0.05$. $^{***}p \leq 0.01$.

profits between Baseline and Competition ($p = 0.001$) and between Fourth Party and Competition ($p = 0.001$) are statistically significant, whereas the difference between Baseline and Fourth Party ($p = 0.481$) is not statistically significant.[15] In column (4) of Table 5, we display the results of a random effects regression with the period-wise earnings of public good players as the dependent variable. We use again Baseline as the benchmark. The results of this regression and a χ^2-test between *Fourth Party* and *Competition* support our previous finding that earnings were highest in Competition. Furthermore, we observe that 'period' has a significant negative effect on earnings.

Result 4. *Contributions are higher in* Fourth Party *and* Competition *than in* Baseline*, and are partly driven by conditional cooperation. The* Competition *treatment generated the highest earnings for public good players.*

5 Concluding remarks

In this work, we compared three different environments governing norm enforcement by third parties and thereby possibly mitigating free-rider problems. In our laboratory ex-

[15]The Wilcoxon tests are based on means of individual profit averages (n = 36 in all treatments).

periment, third parties observed contributions in a linear public goods game with three players, but were not otherwise affected by the outcomes in the public goods game. We then compared punishment decisions by third party players (i) without any additional incentive layer, (ii) when their actions are observed and potentially penalized by a fourth party, and (iii) when their punishment proposals compete against those of another third-party candidate. Our evidence suggests that the democratic election of the norm enforcement authority is utility-maximizing among these three options. This institutional set-up endogenously led to focused sanctions against the worst contributor. Experimental studies where players could directly choose between automatic enforcement of this kind and peer-to-peer punishment found it to be popular and enough to maintain cooperation (see Andreoni and Gee 2012; Kamijo et al. 2014). Our findings also complement results from the second-party punishment literature showing that voting allows subjects to agree on more efficient punishment schemes (see, e.g., Ertan et al. 2009; Putterman et al. 2011). In contrast, we found the effects of having an independent fourth party player to be more ambiguous. While public good players were more cooperative when a fourth party was present, this came at a considerably higher cost of punishment.

Lastly, we note that the effect of institutional environments on third party behavior offers several avenues for future research. For example, it is an open question to which extent the positive effects of electing a sanctioning authority that we found here would hold up when more than two candidates compete. Would increased competition give rise to a 'race to the bottom' in terms of imposed punishment? Another field of interest which we left aside concerns the endogenous selection of an incentive environment for third parties. Our findings here suggest that this might well depend on the relative importance that decision-makers attach to impartiality vs. efficiency.

Declaration of competing interest

None.

References

Andreoni J., and L.K. Gee (2012). Gun for hire: delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics 96*, 1036–1046.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.

Bock, O., I. Baetge, and A. Nicklisch (2014) hroot: Hamburg registration and organization online tool. *European Economic Review 71*, 117–120.

Boyd, R., H. Gintis, and S. Bowles (2003). The evolution of altruistic punishment. *PNAS 100*, 3531–3535.

Carpenter, J., and P. Matthews (2012). Norm Enforcement: Anger, Indignation, or Reciprocity? *Journal of the European Economic Association 10*(3), 555–572.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics 14*, 47–83.

Cinyabuguma, M., T. Page, and L. Putterman (2006). Can second-order punishment deter perverse punishment? *Experimental Economics 9*(3), 265–279.

Dal Bò, P., A. Foster, and L. Putterman (2010). Institutions and behavior: experimental evidence on the effects of democracy. *American Economic Review 100*(5), 2205–2229.

Elster, J. (1989). *The Cement of Society.* Cambridge: Cambridge University Press.

Engel, C., and l. Zhurakhovska (2017). You Are In Charge Experimentally Testing the Motivating Power of Holding a Judicial Office. *The Journal of Legal Studies 46*(1), 1–50.

Ertan, A., T. Page, and L. Putterman (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review 53*(5), 495–511.

Fehr, E., and U. Fischbacher (2004). Third-party punishment and social norms. *Evolution and Human Behavior 25*(2), 63–87.

Fehr, E., and K. M. Schmidt (1999). A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics 114*(3), 817–868.

Fehr, E., and T. Williams (2018). Social Norms, Endogenous Sorting and the Culture of Cooperation. *CESifo Working Papers 7003*.

Filiz-Ozbay, E., and E. Ozbay (2014). Effect of an audience in public goods provision. *Experimental Economics 17*, 200–214.

Fischbacher, U. (2007). z-Tree. Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Fu, T., Y. Ji, K. Kamei, and L. Putterman (2017). Punishment can support cooperation even when punishable. *Economics Letters 154*, 84–87.

Gächter, S., L. Gerhards, and D. Nosenzo (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review 97*, 72–86.

Goette, L., D. Huffman, and S. Meier (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review 96*(2), 212–216.

Gross, J., Z. Méder, S. Okamoto-Barth, S., and A. Riedl (2016). Building the Leviathan - voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports 6*(20767).

Herrmann, B., C. Thöni, and S. Gächter (2008). Antisocial punishment across societies. *Science 319*(5868), 1362–1367.

Kamei, K. (2017). Altruistic Norm Enforcement and Decision-Making Format in a Dilemma: Experimental Evidence MPRA Paper 76641, University Library of Munich, Germany.

Kamei, K. (2018). Group size effect and over-punishment in the case of third party enforcement of social norms. *Journal of Economic Behavior & Organization 68*(1), 18–28.

Kamijo, Y., T. Nihonsugi, A. Takeuchi, and Y. Funaki (2014). Sustaining cooperation in social dilemmas: Comparison of centralized punishment institutions. *Games and Economic Behavior 84*, 180–195.

Kerschbamer, R. (2015). The geometry of distributional preferences and a non-parametric identification approach: The equality equivalence test. *European Economic Review 76*, 85–103.

Krupka, E., and R. Weber (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association 11*, 495–524.

Kurzban, R., P. DeScioli, and E. O'Brien (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior 28*(2), 75–84.

Markussen, T., L. Putterman, and J.-R. Tyran (2014). Self-organization for collective action: an experimental study of voting on sanction regimes. *The Review of Economic Studies 81*(1), 301–324.

Martin, J., J. Jordan, D. Rand, F. Cushman (2019). When do we punish people who don't? *Cognition 193*, 104040.

Mathew, S., and R. Boyd (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of National Academy of Sciences 108*(28), 11375–11380.

Nicklisch, A., K. Grechenig, and C. Thöni (2016). Information-sensitive Leviathans. *Journal of Public Economics 144*, 1–13.

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

Pedersen, E.J., W.H.B. McAuliffe, and M.E. McCullough (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General 147*(4), 514–544.

Philippe, A., and A. Ouss (2018). "No Hatred or Malice, Fear or Affection": Media and Sentencing. *Journal of Political Economy 126*(5), 2134–2178.

Putterman, L., Tyran, J.-R., and Kamei, K. (2011). Public goods and voting on formal sanction schemes. *Journal of Public Economics 95*(9–10), 1213–1222.

Traxler, C., and J. Winter (2012). Survey evidence on conditional norm enforcement. *European Journal of Political Economy 28*(3), 390–398.

Tyran, J.-R., and L.P. Feld (2006). Achieving compliance when legal sanctions are non-deterrent. *The Scandinavian Journal of Economics 108*(1), 135–56.

Xiao, E., and D. Houser (2005). Emotion expression in human punishment behavior. *Proceedings of National Academy of Sciences 102*(20), 7398–7401.

Appendix

A Instructions (english translations)

Welcome to the experiment and thank you for your participation.[16]

In this experiment, all participants have to make decisions. Your payoff will be determined by your own decisions and the decisions of the other participants. You will be paid individually, privately, and in cash after the experiment. During the experiment, we will use the term "points" instead of Euros. Points will be converted into Euros as follows: 100 points = 3 Euros.

Please take your time reading the instructions and making your decisions. You are not able to influence the duration of the experiment by rushing through your decisions, because you always have to wait until the remaining participants have reached their decisions. The experiment is completely anonymous. At no time during the experiment nor afterwards will the other participants know which role you were assigned to and how much you have earned.

If you have any questions please raise your hand. One of the experimenters will come to you and answer your questions privately. Following these rules is very important. Otherwise the results of this experiment will be worthless.

The experiment consists of two parts. Each part will be explained separately. In each part, you can earn money. Your final payoff is calculated as the sum of the payoffs of Part 1 and Part 2. The expected duration of the experiment is 75 minutes.

A.1 Part 1

In the first part, we will ask you to make 10 decisions. In each decision, you are assigned to a group with another participant, who is called "passive agent". Your decision as an "active decision maker" and the decision of the passive agent are made anonymously. In each of the 10 decisions, the passive agent is a different randomly chosen participant. In all decisions, you have to choose between a left and a right option. The options are payoff distributions, meaning that both options are associated with a payoff for you and for the passive agent. An example is given in Table A1.

Example: The left option in the second row in Table A1 is: You 45 points, "passive agent" 65 points. The right option in the second row is: You 50 points, "passive agent" 50 points. If you picked the left option in the second row and the situation is randomly selected as payoff relevant, you would get a payoff of 45 points and the "passive agent" 65 points. (Note that you will see other numbers during the experiment.)

We ask you to decide for each of the 10 decisions between the left and right options. The 10 decisions will be presented in two blocks of 5 decisions each. Please compare row by row the left and right options and decide on your preferred distribution for each row. You can make your

[16]The original instructions were in German. This is an example for the Baseline Treatment. The instructions for the other treatments are available on request.

decision by clicking on the left or right button.

Calculation of your payoff in Part 1

Your payoff from Part 1 results from two partial payoffs. The first partial payoff results from the situation in which you were the active decision maker. At the end of Part 1, the program will randomly select 1 of the 10 decisions. For this decision situation, your decision between left and right will determine the payoff for yourself and the passive agent.

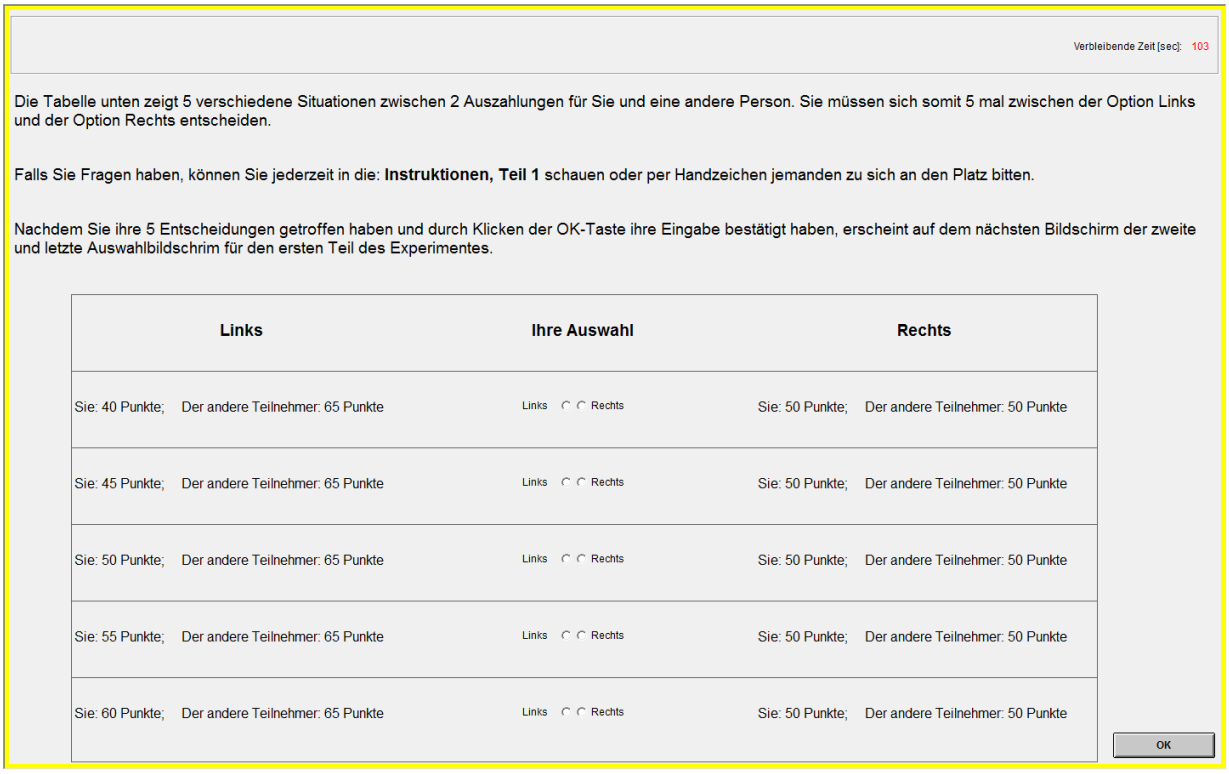
The second partial payoff results from the situation in which you were the passive agent. Following the same procedure as mentioned above, another participant is randomly selected and determines with her chosen left-right-decision your payoff in the role of being the passive agent. We make sure that no two participants are in a reciprocal relation of being an active decision maker and a passive agent for the same person.

Your total payoff from the first part of the experiment is calculated by adding the payoffs from the situations in which you were the active decision maker and the passive agent.

If you have any questions, please raise your hand. One of the supervisors will come to you and answer your questions.

If you do not have further questions, please start and make your decisions between the left and right options.

Fig. A1. Decision screen in Part 1



A.2 Part 2

The second part is played for 20 periods, i.e., the game is repeated 20 times in a row. At the beginning of Part 2, you are randomly assigned to a role (*A-Player* or *B-Player*). In total, there are 18 players of type A and 6 players of type B. Your role stays the same for the entire 20 periods.

At the beginning of every round, all A-Players are randomly assigned to a group that consists of 3 A-Players each (Players A1, A2 and A3). Furthermore, all A-Players receive a budget of 20 points at the beginning of each round. A-Players have to decide how many points of their budget they are willing to contribute to a "public project" that is implemented within their group. Each A-Player can contribute 0, 4, 8, 12, 16 or 20 points.

In every period, one B-Player is randomly assigned to each of the six groups. All B-Players receive a budget of 30 points at the beginning of each round. After the decision of the A-Players, the B-Players are given information about the individual contributions of the players A1, A2 and A3 in his group to the public project. Then, each B-Player can assign punishment points out of his budget to each A-Player in his group. One punishment point reduces the payoff of an A-Player by two points and costs the B-Player one point himself.

Payoff of A-Players in each round

The payoffs of players A1, A2 and A3 depend on the individual contribution to the public project, the contribution of the two other group members and the punishment points that were assigned by the B-Player of the group. Each punishment reduces the payoff of A-Players by two points. The payoff for A-players is calculated using the following formula (players A1, A2 and A3 are A=1,2,3):

*Payoff of A in each round = budget of A - contribution of A + 0.5 * (all contributions within the group) - punishment points of B-Player * 2*

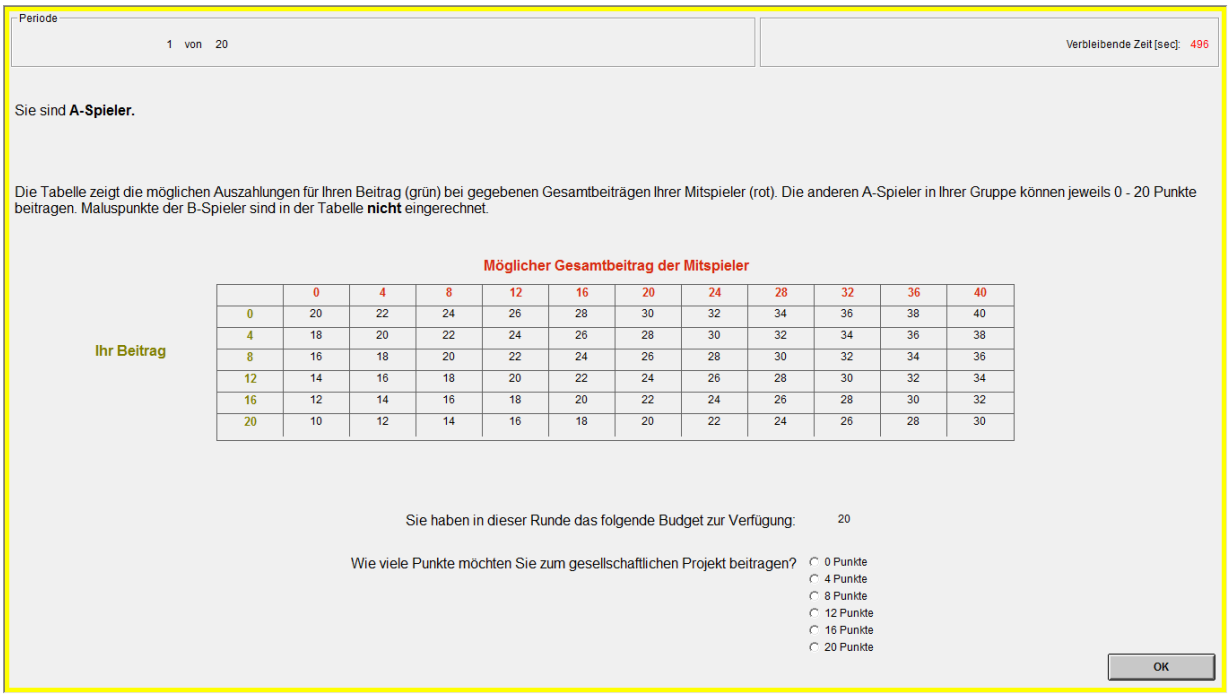
You cannot receive a negative payoff. If the formula yields a negative amount, your payoff is 0 points. Table A2 displays how your own contribution to the public project and the contribution of the other two group members affect your payoff. The table will also appear on the screen when you have to make your decisions during the experiment.

Explanation of table A2 and examples

The table shows possible payoffs for your own contribution to the public project (green) given the contributions of the other group members combined (red). Each group member can contribute 0, 4, 8, 12, 16 or 20 points. The smallest possible contribution of the two group members is 0 points (both players contribute 0 points) and the largest possible contribution of the two group members is 40 points (both players contribute 20 points). Therefore, the table shows all possible contribution combinations of the three A-Players and the individual payoff for a given combination. The payoff is calculated using the formula above. **Punishment points are not included in Table A2**.

Example 1: If the A-Player contributed 8 points and the other group members contributed 16 points combined, then the A-Player would get 24 points (calculation: $20-8+0.5*(8+16)=24$).

Fig. A2. Decision screen in Part 2



The payoff of the A-Player in the current period equals 24 points minus the assigned punishment points of the B-Player in his group times 2. If the B-Player assigned 3 punishment points, for example, the payoff of the A-Player in this round would equal 18 points ($24 - 3 * 2 = 18$).

Example 2: If the A-Player contributed 4 points and the other group members contributed 32 points combined, then the A-Player gets 34 points (calculation: $20 - 4 + 0.5 * (4 + 32) = 34$). If the B-Player assigned 2 punishment points, the payoff of the A-Player in this round would equal 30 points ($34 - 2 * 2 = 30$).

Payoff of A-Players in each round

For the payoff of a B-Player, only the punishment points that he assigned to the three A-Players in his group are relevant. The payoff is calculated as follows:

Payoff of B in each round = budget of B - assigned punishment points to A1, A2 and A3

Example: If the B-Player assigned 2 punishment points to player A1, 2 punishment points to player A2 and 5 punishment points to player A3, then the payoff of the B-Player in this round would equal 20 points (calculation: $30 - 2 - 3 - 5 = 20$).

Final payoff in Part 2

After all A- and B-Players have made their decision in each round, the payoffs for each round are calculated. At the end of each round, you receive information on how many points you earned.

Your final payoff of Part 2 is calculated as the sum of the payoffs of all 20 rounds. At the end of Part 2, each participant receives information on his total payoff in points and the converted total payoff in Euros.

There will be control question before the second part of the experiment starts. The second part only commences if all participants have correctly answered all control questions.

B Additional figures and regressions

Table B1.
Assigned punishment (group level)

Variable	(1) Punishment decision	(2) Punishment decision	(3) Punishment level	(4) Punishment level
Fourth Party	2.002***	1.941***	0.654	0.611
	(0.656)	(0.737)	(0.592)	(0.656)
Competition	0.899*	0.826	0.460	0.403
	(0.519)	(0.588)	(0.597)	(0.654)
SD of group contributions	0.145***	0.166***	0.136***	0.161***
	(0.029)	(0.034)	(0.027)	(0.028)
Sum of group contributions	-0.014*	-0.010	-0.016***	-0.015***
	(0.007)	(0.006)	(0.005)	(0.005)
SD of group contributions $(t-1)$		-0.035		0.018
		(0.022)		(0.014)
Sum of group contributions $(t-1)$		0.004		0.002
		(0.008)		(0.005)
Period		0.017		0.027*
		(0.013)		(0.015)
WTP^a		1.077		-0.072
		(0.794)		(1.065)
WTP^d		0.136		0.033
		(0.592)		(0.621)
Constant	-0.826*	-1.465*	1.048*	0.506
	(0.500)	(0.774)	(0.551)	(0.863)
Observations	960	912	609	576
Wald-χ^2	31.05***	34.86***	37.79***	55.19***

Notes: (1) and (2): Random-effects probit regressions where the dependent variable takes a value of 1 if the B-Player assigned punishment points to at least one of the A-Players in the group and 0 otherwise. (3) and (4): Random-effects regressions where the dependent variable is the number of average punishment points a B-Player assigned to the three A-Players in his group; the dependent variable in (3) and (4) is truncated at zero. Standard errors, clustered at the individual level in all 4 regressions, in parentheses. Deviation: Individual contribution minus average contribution within the group. SD: Standard deviation. $^*p \leq 0.10$. $^{**}p \leq 0.05$. $^{***}p \leq 0.01$.

Fig. B1. Number of punished players for third parties in Fourth Party by assessment (fractions)

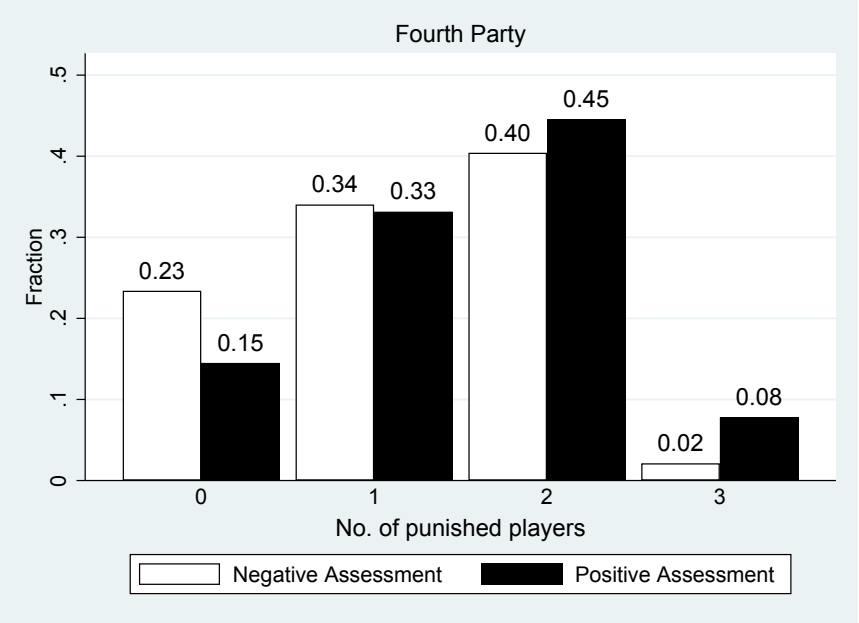


Table B2.
Success of third parties in Fourth Party and Competition (2)

Variable	(1)	(2)	(3)	(4)
	Fourth Party		Competition	
Punishment of norm deviators	-0.085		-0.273	
	(0.442)		(0.322)	
Average punishment within group	0.703***	0.624**	-0.407***	-0.400***
	(0.224)	(0.287)	(0.157)	(0.118)
One player punished (lowest contr.)	0.069	0.068	1.167***	1.016***
	(0.251)	(0.236)	(0.276)	(0.215)
Deviation to punishment profile		0.019		-0.021
		(0.026)		(0.025)
Constant	0.536	0.470	0.113	0.076
	(0.543)	(0.422)	(0.181)	(0.191)
Observations	240	240	480	480
Wald-χ^2	17.258***	10.419**	37.951***	32.306***

Notes: Random-effects logit regressions. Dependent variable: Positive assessment of third party by fourth party in (1) and (2) ($yes = 1/no = 0$) / Electoral success, i.e., third party received majority of votes from public good players in (3) and (4) ($yes = 1/no = 0$). Punishment of norm deviators: third party punished all public good players who contributed less than group average ($yes = 1/no = 0$). One player punished (lowest contr.): Punishment of only one player who is also the lowest contributor to the public good within a group ($yes = 1/no = 0$). Deviation to punishment profile: Sum of squared deviation to the following punishment profile: no punishment for third players with contributions over 8; 1 punishment point for third players with contributions of 8; 2 punishment points for third players with contributions of 4; 3 punishment points for third players with contributions of 0. $^*p \leq 0.10$. $^{**}p \leq 0.05$. $^{***}p \leq 0.01$.

Economics Working Papers

2020-01: Nikolaj Kirkeby Niebuhr: Managerial Overconfidence and Self-Reported Success

2020-02: Tine L. Mundbjerg Eriksen, Amanda Gaulke, Niels Skipper and Jannet Svensson: The Impact of Childhood Health Shocks on Parental Labor Supply

2020-03: Anna Piil Damm, Helena Skyt Nielsen, Elena Mattana and Benedicte Rouland: Effects of Busing on Test Scores and the Wellbeing of Bilingual Pupils: Resources Matter

2020-04: Jesper Bagger, Francois Fontaine, Manolis Galenianos and Ija Trapeznikova: Vacancies, Employment Outcomes and Firm Growth: Evidence from Denmark

2020-05: Giovanni Pellegrino: Uncertainty and Monetary Policy in the US: A Journey into Non-Linear Territory

2020-06: Francesco Fallucchi and Daniele Nosenzo: The Coordinating Power of Social Norms

2020-07: Mette T. Damgaard: A decade of nudging: What have we learned?

2020-08: Erland Hejn Nielsen and Steen Nielsen: Preparing students for careers using business analytics and data-driven decision making

2020-09 Steen Nielsen: Management accounting and the idea of machine learning

2020-10 Qazi Haque, Nicolas Groshenny and Mark Weder: Do We Really Know that U.S. Monetary Policy was Destabilizing in the 1970s?

2020-11 Giovanni Pellegrino, Efrem Castelnuovo and Giovanni Caggiano: Uncertainty and Monetary Policy during Extreme Events

2020-12 Giovanni Pellegrino, Federico Ravenna and Gabriel Züllig: The Impact of Pessimistic Expectations on the Effects of COVID-19-Induced Uncertainty in the Euro Area

2020-13 Anna Folke Larsen, Afonso Saraiva Câmara Leme and Marianne Simonsen: Pupil Well-being in Danish Primary and Lower Secondary Schools

2020-14 Johannes Schünemann, Holger Strulik and Timo Trimborn: Anticipation of Deteriorating Health and Information Avoidance

2020-15 Jan Philipp Krügel and Nicola Maaser: Cooperation and Norm-Enforcement under Impartial vs. Competitive Sanctions