Economics Working Papers

2020-15

Cooperation and Norm-Enforcement under Impartial vs. Competitive Sanctions

Jan Philipp Krügel and Nicola Maaser

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

January 5, 2022

Abstract

We investigate, both theoretically and through a laboratory experiment, how different forms of higher-order punishment affect third party behavior and the level of cooperation in a public goods game. 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 one treatment, the third party is evaluated by another uninvolved individual ("fourth party"); in the other, the third party faces competition by another potential third party punisher. We find that third parties punish free-riders more severely if they have to fear negative payoff consequences for themselves. Our results point to substantial qualitative differences between the institutional arrangements: When the third party is under scrutiny of a fourth party, punishment is high compared to the other treatments, while free-riding is at its lowest. By contrast, competition between two third party candidates leads to strategic and partial punishment.

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

JEL codes: C92; D02; H41

We have benefited from comments by David Levine, Stefan Traub and Fabian Paetzel, and feedback on presentations 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

Numerous studies have documented the willingness of individuals to punish norm violators even though they were not directly harmed by the violation (e.g., Fehr and Fischbacher 2004; Mathew and Boyd 2011; Traxler and Winter 2012). Given the importance of third party punishment for promoting and enforcing cooperative and norm-following behavior, societies have developed additional layers to properly incentivize third parties and to hold them accountable.

One common approach is to introduce a higher-level authority or supervisor who monitors those supposed to act as third parties. For example, managers in firms can and "ought to" implement sanctions against transgressive employees, but are themselves held to account by their superiors when failing to act or misjudging. In other situations, the media or the broader community might play the role of higher-level authority, e.g., a villager who fails to join sanctions against a norm violator can get ostracized himself by the community (see, e.g., Elster 1989, p.127). Often sanctions against third parties, e.g., in form of a written warning or unfavorable media coverage, can remain a threat that is not carried out.

An alternative institutional approach to achieve "good" third-party enforcement is to select, and deselect, 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.

The present paper uses theory and a laboratory experiment to answer two natural questions: How does the behavior of third parties differ under these two frameworks, higher-level scrutiny and competitive election? Which is more effective in maintaining cooperation among unrelated individuals?

Our experiment builds on a public goods game, a standard model of social dilemmas often faced by communities. Importantly, previous research, e.g., Cubitt et al. (2011), has established that free-riding on others in this situation is widely seen as a moral problem. A third party observes contributions in the public good game and can choose to punish individual players at a cost to himself. In addition to this Baseline game, we vary the third party's incentives as follows: (i) another uninvolved outside player, referred to as Fourth party, evaluates and possibly sanctions the third party. (ii) in Competition, two third party candidates compete via their punishment proposals to be selected as the third party by the participants of the public good game. The negative payoff consequences for the third party (candidate) when others – the fourth party or the majority of public good players – disapprove of their punishment behavior are identical.

A main finding is that third parties punish most harshly and target the greatest number of public good players when acting under fourth party scrutiny, even though contributions are at the highest level in this treatment. By contrast, electoral competition leads third parties to direct severe punishment strategically on a subset of public good players who contributed least. Second, our experimental results show that, while the earnings of public good players improve in both treatments compared to the Baseline, Competition did best in this respect, as it combined reasonably high contributions with low average punishment.

One example highlighting the importance of our research questions is the debate in the U.S. about whether law enforcement officers are better appointed or elected. Recently, the residents of King County, Washington (i.e., the Seattle area), passed a ballot measure to replace elected sheriffs with an appointed position that is directly overseen by local administrators (Wissel 2020; Kunkler 2020).[1] Besides being formally responsible for law-enforcement, sheriffs also act as third parties to their districts, e.g., when engaging in informal direct interaction and intervention with citizens (Baldi and LaFrance 2013, p. 149). Yet, surprisingly little is known about the impact of appointing versus electing on how sheriffs exercise their discretion and on cooperative behavior in the community.

Still, to the best of our knowledge, different mechanisms to motivate and control a third party have not been investigated so far using the controlled environment of a laboratory experiment. By providing such a study, our work is related to studies such as Kurzban et al. (2007) and Kamei (2017, 2021) that examine how third parties' behavior is affected when their decisions to punish will be known by others (see Filiz-Ozbay and Ozbay 2014 on audience effects). A key difference with this literature is that in our experiment the third party is not only being watched, but her decisions are subject to explicit evaluation involving potential payoff consequences.[2]

While higher-order sanctions have been studied in the context of second-party punishment, i.e., the question whether non-punishers are punished by their peers (Cinyabuguma et al. 2006; Fu et al. 2017), sanctions against third parties are relatively unexplored. An exception is Martin et al. (2019). Their study compares how often higher-order punishment was targeted against second and third parties, who were the victim of and observed, respectively, a theft and who then *failed* to punish the perpetrator. They found that higher order sanctions were more common against non-punishing observers than against non-punishing victims. This finding indicates that third party punishment has a more normative character – an observer who does not react or reacts too much or too little to an observed norm violation will often be seen as violating normative standards of behavior himself. A key difference to our study is that Martin et al. (2019) focus on the psychological mechanisms underlying higher-order punishment (which they suggest are similar to third-party punishment), whereas we focus on the effect of institutional arrangements.

Finally, our work also relates to the substantial body of research about elected and appointed officials, including studies about judges, prosecutors, regulators, city treasurers or school officials (e.g., Besley and Coate 2003; Huber and Gordon 2004; Partridge and Sass 2011; Whalley 2013; Bandyopadhyay and McCannon 2014). A common finding is that electoral incentives lead to manipulation of outcomes, in particular ahead of elections, in order to please voters. For example, Lim (2013) showed that elected judges with more conser-

[1] The sheriff's office also switched from an elected to an appointed position or the other way around in counties in Connecticut, Florida and Oregon (see Zoorob 2019).

[2] A second difference concerns the game being played. The studies mentioned above use a trust game and a prisoner's dilemma, whereas we consider a public goods game.

vative constituents are more likely to impose harsher sentences.[3] By contrast, appointed officials subject to independent review have been sometimes associated with less biased decision-making (e.g., Hainmueller and Hangartner 2019). However, using observational data to study our research questions is challenging due to the many confounding factors, selection problems and lack of suitable data that would allow to compare norm-enforcing behavior and outcomes under the different sets of incentives.

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 proofs and additional materials in four appendices.

2 Theoretical framework and hypotheses

2.1 Model

The Baseline game. Consider a group $S = \{1, 2, \dots, m\}$ where we assume the group size, $|S|$, to be odd to avoid complications later on. Each individual $i \in S$ can use part of his personal endowment to make a voluntary contribution c_i to a public good that benefits all group members, and keep the rest. We refer to the members of group S as *public good players*.

We assume that each player i is motivated by both material self-interest, which gives them incentives to free-ride on the contributions of others, and a preference for adhering to a social norm (Krupka and Weber 2013; Gächter et al. 2017). Building on d'Adda et al. (2020), we think of a contribution norm N^I as containing two elements: first, individuals' idea about the-right-thing-to-do, $r \in \mathbb{R}_+$, which is a primitive notion in our analysis,[4] and second, a player's expectation of how much others contribute, $\mathbf{E}^I(c)$. Specifically, N^I is the value of a function that computes a weighted average of these two elements:

$$N^I = (1 - \beta)r + \beta \mathbf{E}^I(c). \tag{1}$$

The weights $1 - \beta$ and β ($\beta \in (0, 1)$) measure, respectively, the importance attached to individual values and to conformity with others' contribution behavior. Since different punishment institutions may affect expected contribution behavior and thus the applicable norm, we write $\mathbf{E}^I(c)$ and N^I to emphasize the dependence on the institutional setting. We consider $I \in$ Base, FP, Comp, referring to our Baseline, Fourth party and Competition models of third party punishment.

An outside authority, or *third party* (T), who does not take part in the public good game (PGG), observes all contributions and experiences a disutility when the normative

[3]Relatedly, Huber and Gordon (2004) showed that, while elected trial judges sentence more harshly close to a contested reelection, they are more lenient without electoral pressure, when they arguably can follow more their own preferences (ibid., p. 250).

[4]To focus on the effect of different institutional rules on players' behavior, we abstract from differences in individual values, i.e., we assume that r is identical for all individuals.

standard N^I is violated. This disutility could reflect, for example, inequity aversion (Fehr and Schmidt 1999), or anger towards violators (Jordan et al. 2016). Punishment is costly to the punished, but also to the third party. T chooses a punishment vector $\mathbf{p} = (p_1, \ldots, p_m)$ with the goal to minimize the loss function

$$\mathcal{L}_T = \kappa \sum_{i \in S} p_i + \sum_{i \in S} \frac{(N^I - c_i - p_i)^2}{2\theta_T}, \tag{2}$$

where $p_i \geq 0$ is the punishment directed at public good player $i \in S$, and $\kappa \in (0, 1)$ captures T's cost per unit of administered punishment. Punishment enters negatively into the norm-related terms of T's loss function (2), capturing our assumption that the loss from a norm violation is "healed" to some extent when the violator is punished (Xiao and Houser 2005).

Similarly, each public good player strikes a balance between material and normative normative concerns when deciding how much to contribute to the public good. Player i forms an expectation of the punishment he will face when contributing c_i and the vector of all contributions is $\mathbf{c} = (c_1, \ldots, c_m)$. Each player $i \in S$ minimizes the loss function

$$\mathcal{L}_i = c_i + \lambda \mathbf{E}(p_i|\mathbf{c}) + \frac{(N^I - c_i)^2}{2\theta_i} + \sum_{\substack{j \neq i \\ j \in S}} \frac{(N^I - c_j - p_j)^2}{2\theta_i}\,. \tag{3}$$

The first two terms capture i's loss stemming from his contribution and expected punishment, where parameter $\lambda \in (0, 1)$ measures the utility cost from each unit of punishment i receives. The third and fourth terms capture i's concern for norm-appropriate behavior by self and other group members.

Players' loss (2) and (3) depends on their privately known type θ_T and θ_i, respectively, which captures the steepness of the trade-off between an individual's material and normative interests. The closer θ is to zero, the more importance the individual attaches to norm-compliant behavior. Types are drawn independently from a uniform distribution on $[0, 1]$, and this is common knowledge.[5] Uncertainty in the model is solely with respect to this type, i.e, other parameters are common knowledge.

The Fourth party and Competition games. The FOURTH PARTY game additionally includes another outsider, referred to as the *fourth party*, F, who has the same normative concerns as the public good players and T, but is not in a position to punish public good players herself. F can, however, disapprove T's decisions at no cost to herself.[6] Disapproval

[5] We assume a uniform distribution of types for computational ease; the results below generalize to other continuous distributions whose density is positive on the support.

[6] In real life, individuals who were expected to act as a third-party punisher are sometimes punished symbolically when failing to live up to expectations. Punishment in the form of gossip, verbal reproach or unfavorable coverage in the media is widely seen as costless to the punisher(s) (see Guala 2012). Feinberg et al. (2012) show that gossip is motivated by the same negative affective response that underpins material punishment.

adds an amount D to T's loss, so that loss function (2) becomes

$$\mathcal{L}_T = \kappa \sum_{i \in S} p_i + \sum_{i \in S} \frac{(N^I - c_i - p_i)^2}{2\theta_T} + D \cdot 1_{\{\text{disapproval}\}} \,. \tag{4}$$

We assume that D is "large", so that F's disapproval is prohibitively damaging for T. This assumption could, e.g., capture the damage to T's reputation or future career from being disciplined by a higher-order authority. Whether F approves T's decisions or not has no consequences for the implementation of the punishments chosen by T.

We assume that F's loss function is given by

$$\mathcal{L}_F = \sum_{i \in S} \min\left\{(N^{\text{FP}} - c_i - p_i)^2, \delta \cdot 1_{\{\text{disapproval}\}}\right\} \,, \tag{5}$$

i.e., punishment of norm-violating public good players imposed by T also reduces F's loss. Note that F does not trade off material versus normative concerns since it is costless to her to express disapproval. The loss function (5) assumes that F applies a tolerance level $\delta \in [0, 1]$ when deciding whether to sanction T.

Finally, in the Competition game, the third party is chosen by the public good players in an election among two candidates, A and B. The candidates announce their punishment proposals after observing the contributions in the public good game, but before votes are cast. A candidate's loss function $\mathcal{L}_j$, $j \in \{A, B\}$, is given by (4), where $1_{\{\text{disapproval}\}}$ now is an indicator function for the event that voters did not elect candidate j. Not being elected adds disutility D to the candidate's loss, e.g., because campaigning unsuccessfully is expensive or because not being elected means not enjoying an attractive salary.

2.2 Equilibrium analysis

We simplify the analysis by assuming that PGG players treat expectation $\mathbf{E}(c)$ as independent from their own contribution, i.e., an individual contribution does not change the norm N^I.[7] We are now ready to state the equilibrium of the Baseline game in the following proposition. All proofs are provided in Appendix A.

Proposition 1. *Suppose that θ has an independent uniform prior on the unit interval. Let $\kappa + \lambda < 1$. In the Perfect Bayesian Equilibrium of the* Baseline *game,*

(i) player i's contribution is

$$c_i^* = \begin{cases} \max\left\{0, N^{\text{Base}} - \theta_i(1-\lambda)\right\} & \text{if } \theta_i > \bar{\theta} \\ \max\left\{0, N^{\text{Base}} - \frac{\kappa}{2}\right\} & \text{if } \frac{\kappa}{2} < \theta_i \leq \bar{\theta} \\ \max\left\{0, N^{\text{Base}} - \theta_i\right\} & \text{if } \theta_i \leq \frac{\kappa}{2} \,, \end{cases} \tag{6}$$

where $\bar{\theta} \stackrel{def}{=} \frac{\kappa}{2(1-\lambda)}$.

[7] This seems realistic in large groups; yet, also in smaller groups, it is not clear to which extent individuals would take this "feedback" channel into account.

(ii) The third party's equilibrium punishment strategy is

$$p_i^* = \max\left(0, N^{\textsc{Base}} - c_i - \theta_T\kappa\right) , \tag{7}$$

i.e., all players who contributed less than $N^{\textsc{Base}} - \theta_T\kappa$ receive punishment.

(iii) The equilibrium norm is

$$N^{\textsc{Base}} = r - \frac{\beta}{1-\beta}\left[\frac{4(1-\lambda)^2 + \kappa^2\lambda}{8(1-\lambda)}\right] . \tag{8}$$

The piecewise linear blue graph in Figure 1a shows the equilibrium contributions as afunction of players' type: Contributions always fall short of norm N^B and decrease in θ_i, that is, a public good player contributes less the more importance he attaches to material concerns relative to normative concerns. Moreover, contributions to the public good are greater (lower), the higher λ (κ).

The third party's optimal punishment policy is to not punish any public good player who contributed at least an amount $N^{\textsc{Base}} - \theta_T\kappa$. Individuals with small values of $\theta_i \leq \frac{\kappa}{2}$ prefer contributions exceeding this amount and receive no punishment, whereas individuals with $\theta_i > \bar{\theta}$ find it optimal to contribute less and incur non-zero punishment. For public good players whose type θ_i falls in the intermediate range it is optimal to contribute exactly the amount that is in expectation sufficient to avoid punishment. Figure 1b illustrates (in blue) who is and who is not punished in equilibrium.

We next turn to the Fourth party game, observing first that minimizing the loss function (5) implies that F will disapprove T's decisions if $\left(N^{\text{FP}} - c_i - p_i\right)^2 > \delta$ for some i. Third parties with type $\theta_T \geq \sqrt{\delta}/\kappa$ need to modify their punishment strategy compared to the situation without a fourth party by choosing the amount of punishment that is just sufficient to avoid disapproval. In contrast, third parties who are intrinsically norm-oriented, with $\theta_T < \sqrt{\delta}/\kappa$, prefer a punishment that is severe enough to satisfy the fourth party anyway.

The following Proposition 2 establishes that in equilibrium the punishment, the contributions, and hence the norm N^{FP} are larger than in the situation without a fourth party.

Proposition 2. *Suppose that θ has an independent uniform prior on the unit interval. Let $\kappa + \lambda < 1$ and $\sqrt{\delta} < \kappa/2$ (non-lenient fourth party). Then, in the Perfect Bayesian Equilibrium of the* Fourth Party *game,*

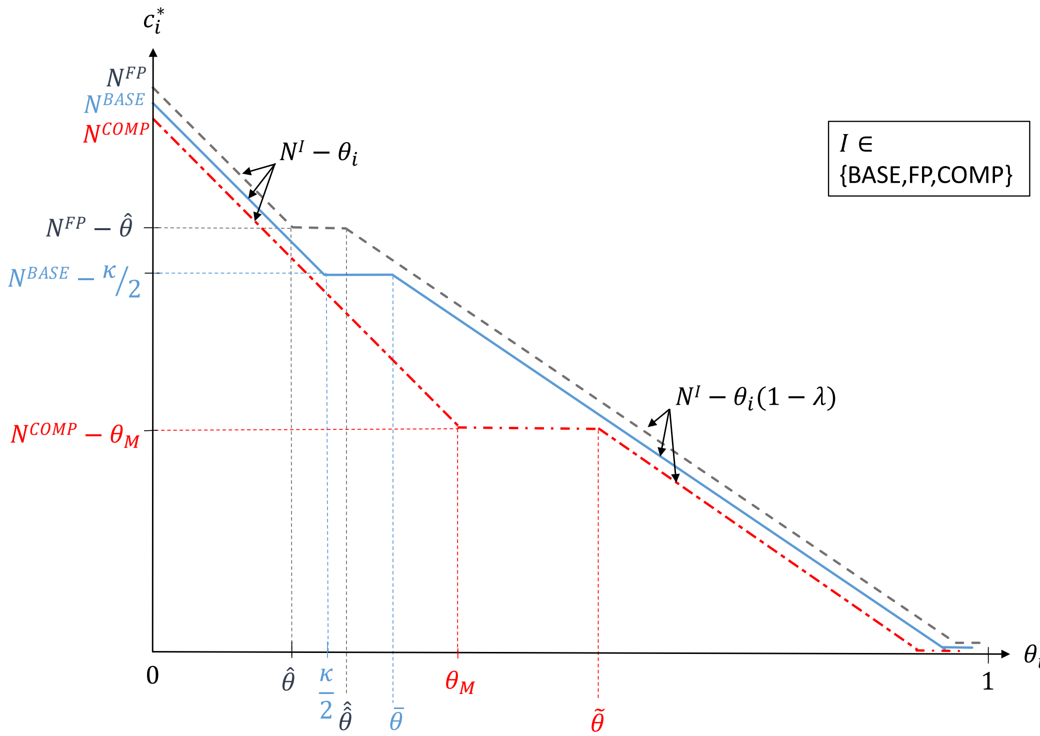
(i) player i's equilibrium contribution is given by

$$c_i^{**} = \begin{cases} \max\left\{0, N^{\text{FP}} - \theta_i(1-\lambda)\right\} & \text{if } \theta_i > \hat{\hat{\theta}} \\ \max\left\{0, N^{\text{FP}} - \sqrt{\delta}\right\} & \text{if } \hat{\theta} < \theta_i \leq \hat{\hat{\theta}} \\ \max\left\{0, N^{\text{FP}} - \theta_i\right\} & \text{if } \theta_i \leq \hat{\theta} , \end{cases} \tag{9}$$

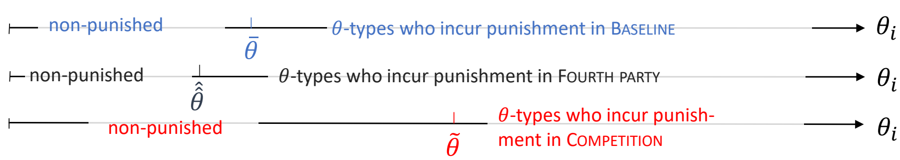
where $\hat{\theta} \overset{def}{=} \sqrt{\delta} - \frac{\delta}{2\kappa}$ and $\hat{\hat{\theta}} \overset{def}{=} \frac{\hat{\theta}}{1-\lambda}$.

Fig. 1. Equilibrium contributions and expected punishment decision in Baseline (blue), Fourth party (grey) and Competition (red).

(a) Equilibrium contributions.



(b) Which types can expect to get punished?



(ii) The third party's equilibrium punishment strategy is

$$p_i^{**} = \max\left\{0, N^{\mathrm{FP}} - c_i - \sqrt{\delta}, N^{\mathrm{FP}} - c_i - \theta_T \kappa\right\} . \tag{10}$$

(iii) The equilibrium norm is

$$N^{\mathrm{FP}} = r - \frac{\beta}{1-\beta}\left[\frac{4(1-\lambda)^2 + 4\delta\lambda}{8(1-\lambda)}\right] + \frac{\beta}{1-\lambda}\frac{\lambda}{1-\lambda}\left[\frac{4\kappa\sqrt{\delta}\delta - \delta^2}{8\kappa^2}\right] . \tag{11}$$

The grey graph in Figure 1a illustrates the equilibrium. As shown in Corollary 1 below, we have $N^{\mathrm{FP}} > N^{\mathrm{Base}}$ for all $\beta > 0$ and $\lambda > 0$, subject to the condition that $\sqrt{\delta} < \kappa/2$. Propositions 1(i) and 2(i) lead to our first hypothesis:

Hypothesis 1 (Amount of contributions). *Public good players' contributions are larger on average if a* Fourth Party *is present compared to* Baseline.

The (ii)-parts of Propositions 1 and 2 imply that public good players who would already have incurred punishment without a fourth party's presence are punished more severely now. Moreover, as shown in Figure 1b, the share of players who incur punishment is greater in Fourth party compared to Baseline. This leads to our second set of predictions:

Hypothesis 2 (How much punishment?).

(a) Punishment per unit of norm violation in the Fourth Party *treatment exceeds that in the* Baseline *treatment.*

(b) The number of punished players in Fourth Party *is greater than that in the* Baseline *treatment.*

In the Competition game, every public good player has two decisions, first choosing his contribution to the public good, and then deciding in favor of one punishment proposal over the other. Clearly, i casts his vote in favor of the candidate whose punishment proposal $\mathbf{p}$ leads to a smaller value of i's loss function (3). If there is no difference, then i chooses randomly, with equal probability for electing either candidate.

Since public good players experience a loss from other group members' norm violations, it is not optimal for a third party candidate to leave norm violations generally unpunished. Rather, electoral concerns cause both candidates to penalize a subset of at most $\frac{|S|-1}{2}$ players who contributed less than the median contributor. This allows a candidate to gain the support of group members who contributed (weakly) more than the median.[8] In equilibrium, the two candidates submit identical punishment proposals and win the election with probability one half. The following proposition summarizes our results:

Proposition 3. *Suppose that players have a common prior about the distribution of θ on $[0,1]$ and let θ_M denote the median of this distribution. In the Perfect Bayesian Equilibrium of the* Competition *game,*

(i) public good player i contributes

$$c_i^{***} = \begin{cases} \max\left\{0, N^{\text{Comp}} - \theta_i(1-\lambda)\right\} & \text{if } \theta_i > \tilde{\theta} \\ \max\left\{0, N^{\text{Comp}} - \theta_M\right\} & \text{if } \theta_M < \theta_i \leq \tilde{\theta} \\ \max\left\{0, N^{\text{Comp}} - \theta_i\right\} & \text{if } \theta_i \leq \theta_M \,, \end{cases} \tag{12}$$

where $\tilde{\theta} \stackrel{def}{=} \frac{\theta_M}{1-\lambda}$;

[8] Under the assumption that the number of public good players, $|S|$, is odd, exactly one median voter generically exists.

(ii) both third party candidates propose identical punishments

$$p_i^{***} = \begin{cases} \max\left\{0, N^{\text{Comp}} - c_i\right\} & \text{if } c_i < med(\mathbf{c}) \\ 0 & \text{if } c_i \geq med(\mathbf{c}), \end{cases} \tag{13}$$

where $med(\mathbf{c})$ is the realized median contribution in the PGG.

(iii) Suppose that θ has an independent uniform prior on the unit interval. Then, the equilibrium norm is

$$N^{\text{Comp}} = r - \frac{\beta}{1-\beta}\left[\frac{(1-\lambda)^2 + \theta_M^2\lambda}{2(1-\lambda)}\right]. \tag{14}$$

(iv) Moreover, if

$$\lambda > \sum_{j \neq i} \frac{N^{\text{Comp}} - c_j - p_j}{\theta_i}, \tag{15}$$

then public good player i prefers the candidate who punishes him less in case that the candidates' proposals are not identical (off the equilibrium path).

Proposition 3(i) implies that the effect of two-candidate competition on contributions is a priori ambiguous. Whether public good players contribute more or less compared to Baseline and Fourth party, and whether the norm is more or less demanding than in these two, depends on the distribution of types, in particular on the location of θ_M. The red graph in Figure 1a depicts a situation where contributions will be lower in Competition than in the other two settings, as is for example the case if types are drawn from a uniform distribution on $[0, 1]$. In this case, we obtain the following result by directly comparing the equilibrium norms (8), (11) and (14):

Corollary 1. *Suppose that players' types are distributed uniformly on $[0, 1]$, $\kappa + \lambda < 1$, and $\sqrt{\delta} < \kappa/2$. Then,*

$$N^{\text{Comp}} < N^{\text{Base}} < N^{\text{FP}}.$$

But for a type distribution with, say, $\theta_M < \bar{\theta}$, equilibrium contributions in Competition would be greater than in the Baseline setting.

Proposition 3(ii) shows that it also depends on the median type θ_M how many public good players get punished in equilibrium. Figure 1b illustrates, in line with panel (a), a situation where the median public good player is relatively uninterested in norm-following behavior. In this situation, we would expect only very few public good players to incur punishment. We formulate our expectations as

Hypothesis 3 (Punishment in Competition).

(a) In the Competition *treatment, both candidates propose to punish at most one public good player, the least contributor.*

(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 to at most one public good player.*

(c) *Conditional on receiving non-zero punishment, players may be punished more or less harshly per unit of norm violation in* COMPETITION *compared to* BASELINE *and* FOURTH PARTY *since $N^{\text{COMP}} - c_i$ can be greater or smaller than $N^{\text{BASE}} - c_i - \theta_T \kappa$ and $N^{\text{FP}} - c_i - \delta$.*

Finally, the condition established in Proposition 3(iv) is more likely to be satisfied, the greater θ_i – and thus the lower i's contributions. For the case that (15) holds, we arrive at the following two additional predictions:

HYPOTHESIS 4 (Elections).

(a) *A public good player does not vote for a candidate who proposes to punish him.*

(b) *A public good player is more likely to vote for the candidate who proposes more (less) punishment the greater (lower) his own contribution.*

3 Experimental design

3.1 The public good game

Our experimental design builds on a standard linear public goods game with three players which is repeated for 20 periods. At the beginning of the first period, all participants are randomly assigned to one of two roles, which stay fixed for the entire 20 periods – public good player ("A-player") or third party ("B-player").[9] In each period, three A-players are randomly matched with one B-player using a stranger matching protocol to avoid individual reputation building.

At the beginning of every period, A-players receive an endowment of 20 points each. They then decide simultaneously and without communication to contribute $c_i \in \{0, 4, 8, 12, 16, 20\}$ to a public good with a marginal per capita return $\alpha = 0.5$, which is implemented in their group. After the PGG players have made their decisions, the third party is informed about the contributions c_1, c_2 and c_3 in her group and can punish A-player i by assigning punishment points p_i to i. One punishment point costs B one point out of his endowment of 30 points, but reduces i's payoff by two points. The monetary payoff π_{A_i} for A-player i is given by[10]

$$\pi_{A_i} = 20 - c_i + 0.5 \sum_{j=1}^{3} c_j - 2p_i \,. \tag{16}$$

[9]We use a neutral framing ("A-player", "B-player" and "public project") in the experimental instructions (see Appendix D).

[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 total punishment points she assigned, i.e.,

$$\pi_{\mathrm{B}}^{Base} = 30 - \sum_{i=1}^{3} p_i. \tag{17}$$

At the end of each period, the participants receive feedback about all decisions of their group members and information about their own payoff in the period. The final payoff for each A-player 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 so that each group consists of three A-players, one B-player and one C-player. Roles stay fixed across 20 periods and groups are randomly rematched in every round. In addition to the two stages described before, a period now involves a third stage, where the C-player receives information about contributions and assigned punishments and can then indicate whether he finds B's decisions "appropriate" or "not appropriate". C-players receive a fixed payoff of 15 euros irrespective of their decision.[11] The payoffs for A-players are still calculated according to (16). The B-player's payoff, however, now depends on C's decision:

$$\pi_{\mathrm{B}}^{Fourth} = \begin{cases} 30 - \sum_{i=1}^{3} p_i & \text{if C agrees} \\ 5 & \text{if C disapproves}\,. \end{cases} \tag{18}$$

The third treatment, COMPETITION, has groups consisting of three A-players as before and two B-player candidates (B_1 and B_2). Both candidates observe A-players' contributions in their group and suggest punishments $(p_1^{B1}, p_3^{B1}, p_3^{B1})$ and $(p_1^{B2}, p_3^{B2}, p_3^{B2})$ to A-players. The A-players are informed about these punishment proposals and indicate their preferred proposal by vote. The proposal which receives the majority of votes is implemented. The payoff for an A-player is calculated by (16), 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 (17); the non-selected B-player receives a payoff of 5 points, i.e.,

$$\pi_{\mathrm{B}}^{Comp} = \begin{cases} 30 - \sum_{i=1}^{3} p_i & \text{if B's proposal received two or three votes} \\ 5 & \text{if B's proposal received zero or one vote}\,. \end{cases} \tag{19}$$

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 conducted the equality

[11] The payoff for C is designed such that it approximately equals the average final payoff of A- and B-players.

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.[12] Subjects completed the Kerschbamer-test first (part 1), before participating in BASELINE, FOURTH PARTY or COMPETITION.[13] As final payoff, each participant received the sum of 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.[14]

The experiment was conducted at the University of Hamburg and involved eight sessions with a total of 228 participants; 48 subjects participated in BASELINE (two sessions, 24 subjects per session), 90 subjects participated in FOURTH PARTY and 90 subjects participated in COMPETITION (both three sessions, 30 subjects per session). 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 75-90 minutes. The highest payoff was €20.52, the lowest payoff €6.93 and the average payoff €15.52. All decisions and payoffs were made in private.

4 Results

We begin with an overview of our main findings before providing a detailed analysis of punishment behavior, contributions, and earnings.

4.1 Overview

To first get an idea of contribution behavior, Figure 2 shows the average contributions to the public good by treatment (left panel) and by treatment and period (right panel). A first result is that FOURTH PARTY generated higher contribution rates than BASELINE (mean 11.57 and 6.59 points, respectively), as predicted in Hypothesis 1. Additionally, we see that COMPETITION had greater contributions (mean 9.54 points) than BASELINE as well.[15] There is a slight, negative contribution trend over time for the BASELINE and

[12]The test provides 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, $WTP^d \in [-0.667, 0.667]$.

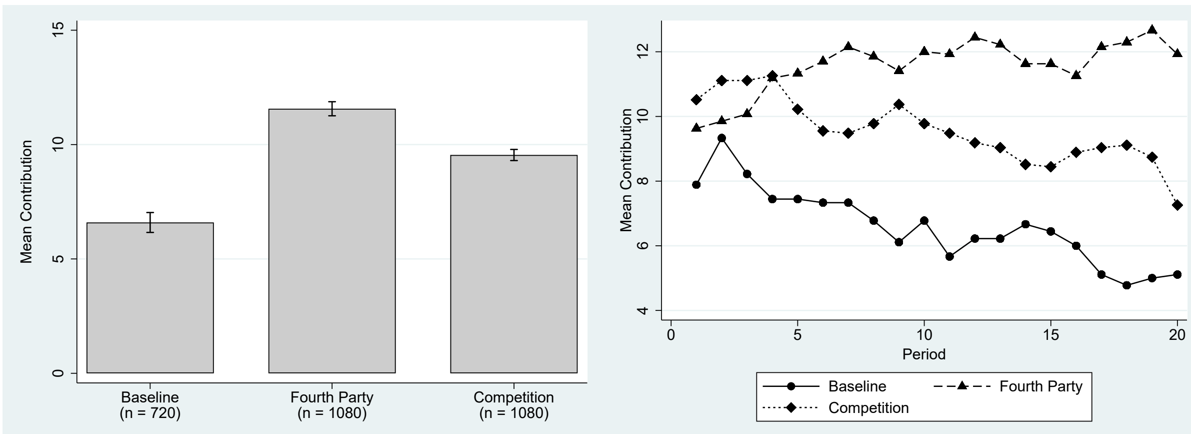
[13]The order of the experimental parts was chosen this way so that participants could familiarize themselves with the experimental environment before completing the main task.

[14]There are established methods for eliciting social norms in experiments (see, e.g., Krupka and Weber 2013). Although social norms are important for contribution and punishment behavior, we see players' beliefs or perceptions about the social norm as not central to our analysis. The reason is that we are primarily interested in how the institutional settings under consideration affect third parties' behavior. In light of this, we chose not to elicit the beliefs about social norms because the elicitation procedure itself may unduly influence behavior and lead to experimenter demand effects.

[15]Wilcoxon ranksum tests with session averages reveal that while the differences between BASELINE and FOURTH PARTY ($p = 0.083$) and between BASELINE and COMPETITION ($p = 0.083$) are significant, the

Competition treatments, that is not observed for Fourth Party. The data replicate the stylized fact from previous PGG experiments that participants initially contribute on average between 40 % and 60 % of their endowment (see Chaudhuri 2011).

Fig. 2. Mean contribution by treatment and over time



Notes: Error bars represent 95% confidence intervals for the mean.

In the second stage of each treatment, third parties could punish PGG players. Figure 3 shows the average punishment points assigned by treatment (left panel) and by treatment and period (right panel). In Competition, there were two third parties in each group. In the left panel of Figure 3, we first incorporate both punishment proposals to calculate the mean for all punishment observations (third bar from the left). The mean punishment was higher in Fourth Party (mean punishment: 1.70 points) than in Competition (mean punishment for all observations: 1.01 points) and Baseline (mean punishment: 1.00 points).[16] The left panel of Figure 3 also shows the mean punishment proposals separately for "rejected" and "accepted" third party candidates in the Competition treatment, i.e., whose proposals respectively received the minority and majority of public good players' votes. We find that on average, rejected third parties (mean punishment: 1.33 points) punished more harshly than accepted third parties (mean punishment: 0.69 points).

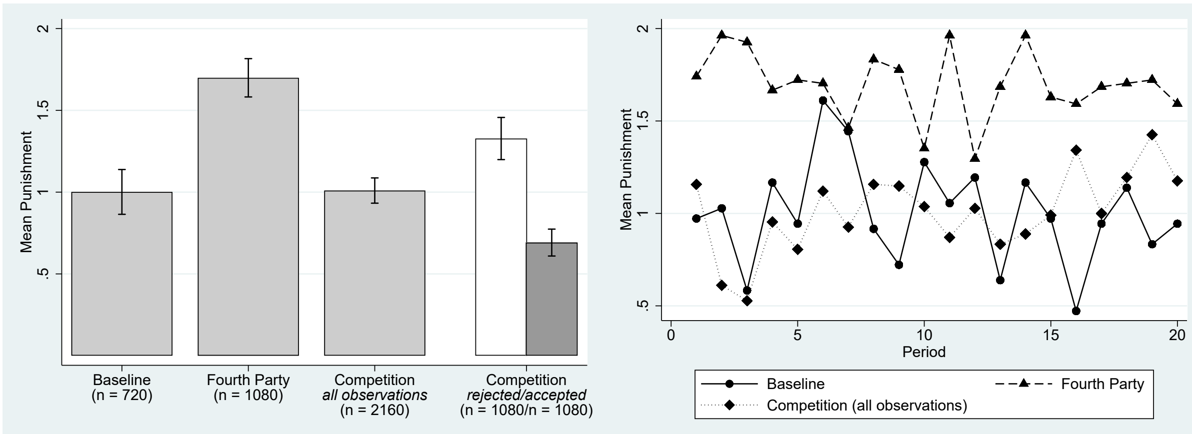
We also find that in 2563 of a total of 3960 individual punishment decisions (64.7%), third party players chose not to punish at all. Many public good experiments with punishment show decreasing levels of punishment over time. Interestingly, there is no clear punishment trend in any of our treatments (see Figure 3, right panel).

We now turn to third party punishment at the group level. Considering all treatments combined, the majority of third parties tended to punish either none or only one of the

difference between Fourth Party and Competition ($p = 0.513$) is not significant.

[16]Wilcoxon ranksum tests with session averages produce the following results: Baseline vs. Fourth Party ($p = 0.248$), Baseline vs. Competition ($p = 1.00$), Fourth Party vs. Competition ($p = 0.049$). However, as we discuss in more detail in Subsection 4.2, the punishment per norm violation that is relevant for comparing the treatments.

Fig. 3. Mean punishment by treatment and over time



Notes: Error bars represent 95% confidence intervals for the mean.

public good players. When only one player was punished, the punishment was almost always applied to the player who contributed the least to the public good – this was true in 97.62% of cases where only one player was punished. We see similar results when two players were punished. In this case, the two players who contributed the least to the public good were punished in 98.58% of the cases.

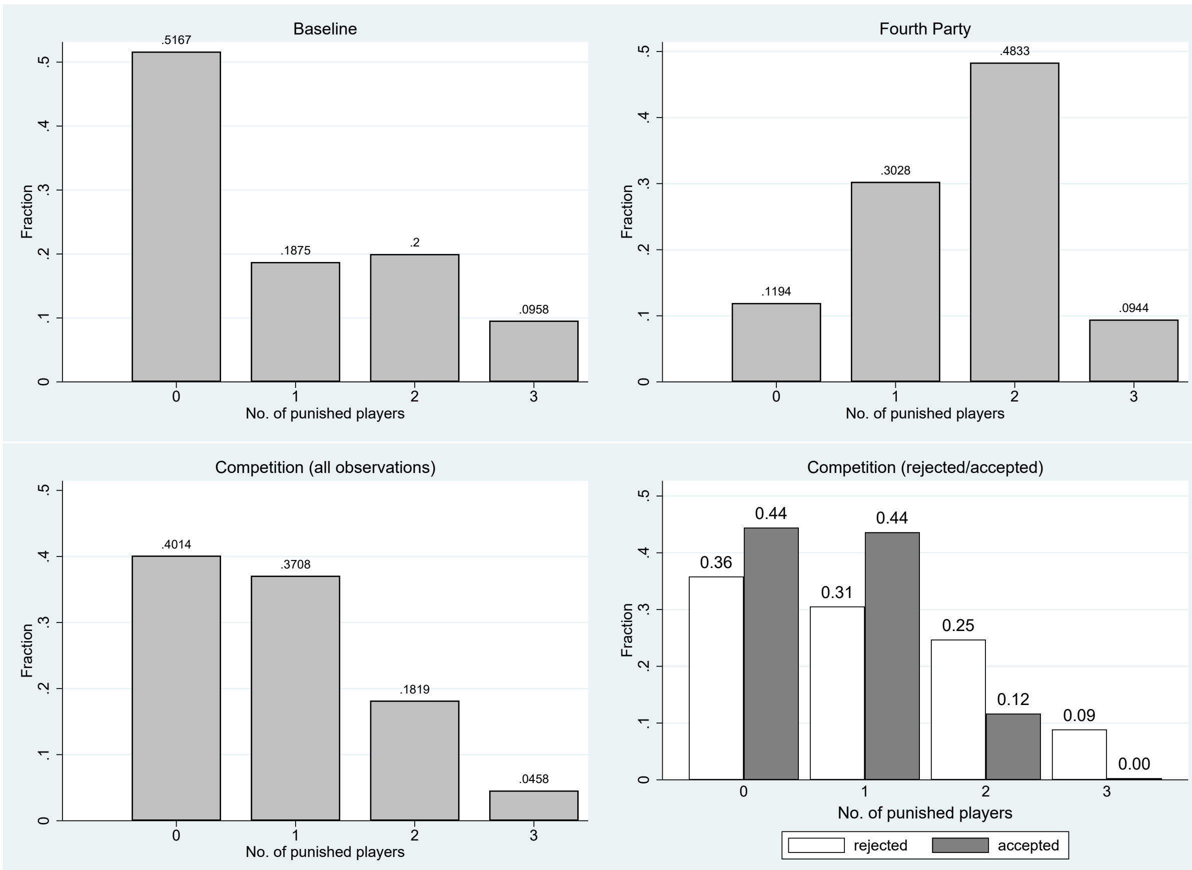
Figure 4 shows how many of the three public good players in a group were punished (in percent, by treatment and pooled over all 20 periods). In Baseline, 51.67% of the time none of the A-players in a group were punished. The picture is very different in Fourth Party, where punishment was not only harsher (cf. Figure 3), but also frequently directed against multiple public good players: Only in 11.94% of cases was no player punished. Punishment of one or two players was much more frequent, with 30.28% and 48.33% of cases, respectively. With respect to the Competition treatment, Figure 4 again shows punishment profiles for "rejected" and "accepted" third party candidates.[17] We find that third parties were most successful when they did not punish any or one public good actor.

Lastly, Figure 5 shows how PGG players' contributions influenced the punishment decisions of the third parties. The plot shows the mean punishment as a function of the deviation (d) of a public good player's contribution from the average contribution in his group.[18] 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 low in all three treatments. Punishment increased significantly when the contribution was negatively

[17]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 C, see Figure C1.

[18]Comparing deviations from the average contribution in the group is both suggested by our formulation (1) of the norm and widely used in studies of PGGs.

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

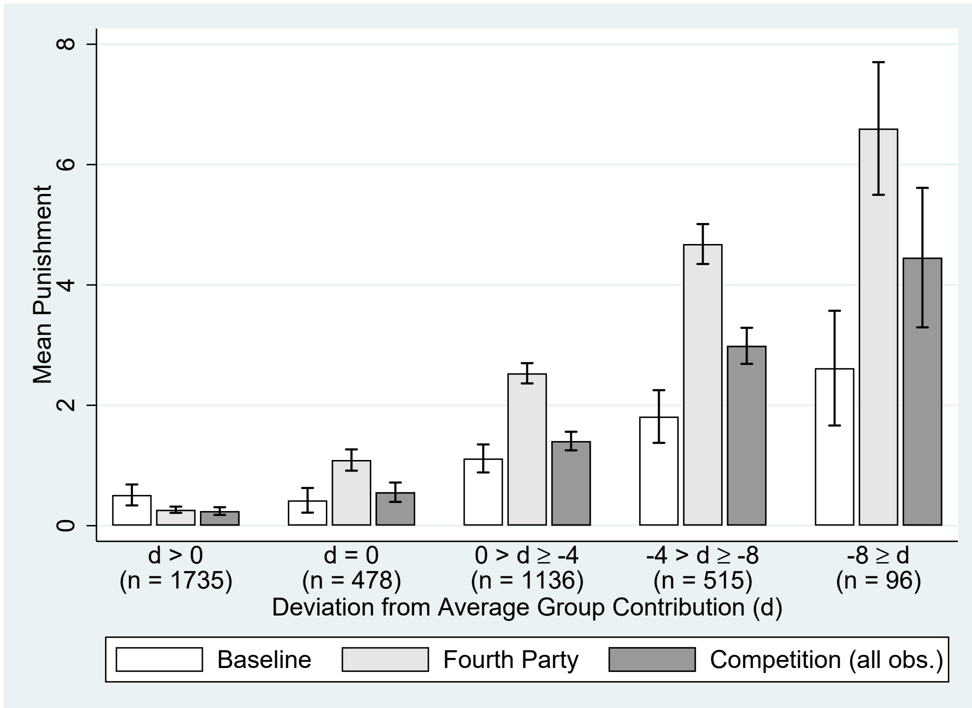


different from the group average. We also observed a stronger effect for larger deviations. The three groups of bars on the right side of Figure 5 indicate that third parties punished negative deviations the most in the Fourth Party treatment.

4.2 Third party behavior

Our main interest is in analyzing the impact of different third-party incentives. We run multiple regressions to identify the key factors that determine the size of the assigned penalty. Since we have 2563 observations of third parties who did not receive a penalty, we analyze the probability that the penalty is nonzero separately from the severity of the penalty. Table 1 shows random effects probit regressions in columns (1) and (2), where the dependent variable takes the value of 1 if the B player punished an A-player and 0 otherwise. Columns (3) and (4) show random effects regressions where the dependent variable is the level of third party punishment (truncated at zero). We use Baseline as our benchmark treatment in all regressions. Standard errors are clustered at the subject level because participants are asked to make repeated decisions with different partners in

Fig. 5. Deviation from average group contribution and individual punishment



Notes: d = individual deviation from the average contribution in the group; Error bars represent 95% confidence intervals for the mean.

the same setting.[19]

The results of the first regression in Table 1 and a Wald-test between *Fourth Party* and *Competition* suggest that third parties chose to punish more often in Fourth Party than in Baseline and Competition. In regression (2), we include two interaction terms, *Fourth Party* x *deviation* and *Competition* x *deviation*. Both terms have a significant and negative impact on punishment. Thus, we find that a treatment difference in the assigned punishment occurs only when the contribution of a public good player negatively deviates from the group average, consistent with the results presented in Figure 5. The interaction terms show that the strongest effect for *deviation* on the punishment decision was present in the Fourth Party treatment. A Wald-test indicates that the difference between the interaction terms *Fourth Party* x *deviation* and *Competition* x *deviation* is statistically significant ($p < 0.01$, see last row of Table 1). We conclude that the punishment decision was more responsive to deviation from the average in Fourth Party. Moreover, 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 pe-

[19]For our main results (Table 1, Table 2, and Table 5), we run additional regressions in which we cluster standard errors at the session level. The results are very similar, both qualitatively and quantitatively (see Appendix B).

riod 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. The regression reveals a significant positive impact of more spread-out contributions on the punishment decision. The other control variables are insignificant.

In regressions (3) and (4), we analyze the chosen level of punishment. The basic treatment effects are insignificant in regression (3). Regression (4) shows that negative deviations of public good players from average group contribution are punished most severely in Fourth Party.[20] In contrast to regression (2), period effects are significant in (4). Generally, we obtain very similar results when considering the punishment level or the punishment decision as dependent variables.[21] We therefore conclude:

Result 1. *If the contribution of a public good player negatively deviates from the group average, third parties punish more severely in* Fourth Party *compared to* Baseline*, thus confirming Hypothesis 2(a). The punishment level per unit of norm violation is higher in* Fourth Party *than in* Competition *(see Hypothesis 3(c)).*

We now turn to group-level analysis by analyzing how many public good actors were punished in each treatment.[22] Table 2 presents logit regressions in which the dependent variable is an indicator variable indicating how many public good actors were punished by the 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 again use the baseline treatment as a benchmark. The regression shows a negative and significant effect of the treatment variable *Fourth Party*. A Wald test shows that the difference between *Fourth Party* and *Competition* is also statistically significant. These results imply that "no punishment" was more common in Baseline and Fourth Party than in Competition. In addition, a high standard deviation in group contributions had a negative and significant effect. That is, when the level of contributions was highly variable within a group, "no punishment" was rare.

In column (2), Table 2, the dependent variable is an indicator equal to one if exactly one player was punished and zero otherwise. We find that a player was punished more often in Competition than in Baseline.[23] However, a Wald test shows that the difference between *Fourth Party* and *Competition* is not statistically significant. Since punishment of actors involved in public goods is a complementary event to "no punishment," it is natural to expect the coefficient on the independent variable to change sign; for example,

[20] A Wald-test confirms that the difference between the interaction terms *Fourth Party* x *deviation* and *Competition* x *deviation* is significant ($p < 0.01$).

[21] The main results of Table 1 are also supported by additional regressions where we cluster standard errors by session (see Appendix B, Table B1).

[22] In addition to our estimates in Table 1, we also ran group-level regressions for the punishment decision and punishment level, which we report in Table C1, Appendix C. In these regressions, we also included lagged variables as additional controls.

[23] Note that the variable *Competition* in Table 2 does not distinguish between successful and unsuccessful third parties.

Table 1.
Estimates for individually assigned punishment

Variable	(1) Punishment decision	(2) Punishment decision	(3) Punishment level	(4) Punishment level
Fourth Party	1.394**	1.182*	0.868	-0.059
	(0.605)	(0.671)	(0.697)	(0.636)
Competition	0.449	0.253	0.865	0.604
	(0.559)	(0.510)	(0.695)	(0.681)
Deviation	-0.258***	-0.146***	-0.333***	-0.135***
	(0.026)	(0.032)	(0.047)	(0.048)
Fourth Party x deviation		-0.285***		-0.307***
		(0.074)		(0.071)
Competition x deviation		-0.092**		-0.126*
		(0.045)		(0.064)
SD of group contributions		0.047*		0.123***
		(0.024)		(0.036)
Period		-0.002		0.040**
		(0.008)		(0.017)
WTP^a		0.721		-0.132
		(0.593)		(0.971)
WTP^d		0.082		-0.259
		(0.370)		(0.699)
Constant	-1.363**	-1.522***	1.354*	0.969
	(0.559)	(0.590)	(0.717)	(0.716)
Observations	3960	3960	1397	1397
Wald-χ^2	103.78***	132.87***	85.03***	206.57***
Fourth Party = *Competition*	$p = 0.005$		$p = 0.996$	
Fourth Party x Deviation = *Competition* x Deviation		$p = 0.0097$		$p = 0.0069$

Notes: (1) and (2) are probit regressions with subject random effects where the dependent variable takes a value of 1 if the B-player punished an A-player and 0 otherwise. (3) and (4) are 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. Robust standard errors, clustered at the subject level in all 4 regressions, in parentheses. Deviation: Individual contribution minus average contribution within the group. SD: Standard deviation. $^*p \leq 0.10$. $^{**}p \leq 0.05$. $^{***}p \leq 0.01$.

Table 2.
Number of punished players in a group: regressions

Variable	(1) no punishment	(2) one player punished	(3) >one player punished
Fourth Party	-3.661***	1.105*	1.901**
	(0.976)	(0.583)	(0.749)
Competition	-1.298	1.503***	-0.042
	(0.849)	(0.573)	(0.742)
SD of group contributions	-0.268***	0.095***	0.121***
	(0.047)	(0.035)	(0.036)
Period	-0.022	0.020	-0.008
	(0.019)	(0.015)	(0.014)
Constant	2.210**	-2.767***	-2.061***
	(0.944)	(0.665)	(0.782)
Observations	1320	1320	1320
Wald-χ^2	53.15***	11.46**	36.70***
Fourth Party = *Competition*	$p = 0.0000$	$p = 0.1598$	$p = 0.0000$

Notes: Logit regressions with subject random effects. Dependent variable: 0 players punished ($yes = 1/no = 0$) / 1 player punished ($yes = 1/no = 0$) / >1 player punished ($yes = 1/no = 0$). Robust standard errors, clustered at the subject level, in parentheses. SD: Standard deviation. *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

the coefficient on "standard deviation of contributions" is now positive instead of negative. In the third regression, the dependent variable is an indicator of ">one penalized player". The third regression and a Wald test between *Fourth Party* and *Competition* confirm that two or three players were significantly more likely to be penalized in Fourth Party than in Baseline and Competition. Overall, these findings support the following result:

Result 2. *In* Fourth Party, *the number of penalized public good players is larger than in* Baseline, *confirming Hypothesis 2(b). Punishment in* Competition *is often strategically targeted only at the largest deviator from the average group contribution, confirming Hypothesis 3(a).*

Next, we take a closer look at the Fourth Party and Competition treatments to see if these key results are confirmed by further analysis. In Fourth Party, there were 360 instances where a fourth party player had to evaluate the third party's punishment decision. As expected in theory, the evaluation was often positive: In 81.67% of the cases, the fourth party agreed with the third party's punishment decision.

Column (1) in Table 3 presents a random effects logit regression in which the dependent variable is a dummy variable ("F agrees") indicating a positive evaluation by the fourth party. 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 the group average, and zero otherwise. The regression shows that this variable does not have a significant effect on the evaluation of fourth parties. In contrast, a higher average punishment within the group does have a positive and significant effect on approval. These results suggest that it is more important to punish extensively, rather than to just punish those who contribute less than the group average or less than half of their endowment.[24]

Column (2) in Table 3 shows a regression that has the electoral success of third parties in Competition as the dependent variable. Just as in the first regression, the variable "punishment of norm deviants" has no significant effect on electoral success in Competition. However, the dummy variable for punishing one public good player has a significant positive effect on electoral success.[25] Moreover, higher average punishment within a group led to significantly lower electoral success. Further illustration is provided by Figure 6, which shows the probability of electoral success when a public good actor is punished compared to other punishment proposals. Consistent with our Hypothesis 3(b), third parties are most successful in winning a majority of votes when they punish one of the three public good actors, especially when the other third party proposes to punish two or three public good actors.[26] 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, in line with the equilibrium stated in Proposition 2(ii).*

In Competition, the public good players chose between the punishment proposals of two candidates. As expected, they almost always (in 98.61% of cases) voted for the candidate who proposed less punishment for them, confirming Hypothesis 4(a). To test Hypothesis 4(b), we run a random-effects logit regression (see Table 4) in which the dependent variable is a dummy variable indicating whether a public good actor voted for the candidate who proposed a higher total penalty ($yes = 1/no = 0$).[27] As explanatory variables, we include own contribution and period. The regression shows that a public good player is more likely to vote for the candidate who proposes more punishment the greater his own contribution was, confirming Hypothesis 4(b).[28]

[24] We conducted additional regressions using an alternative benchmark against which norm violations might be measured. We constructed a hypothetical "ideal" punishment profile and assessed the quality of third party punishment by the sum of the (squared) deviation of actual punishment from this "ideal" profile. The choice of the ideal punishment profile is to some extent arbitrary and we therefore relegate the corresponding results to Appendix C (see Table C2 there).

[25] As mentioned earlier, in the case of punishing only one actor, the one who contributed the least to the public good was almost always punished. The results in columns (1) and (2) are virtually identical when replacing "one actor punished" with a dummy that also captures whether that actor was also the lowest contributor to the public good (see Appendix C, Table C2).

[26] Third parties that do not punish any of the public good actors were also very successful when matched with a competitor that proposed to punish two or three public good actors. Matchups where both third parties decided not to punish ($n = 73$) or where both third parties decided to punish one common good actor ($n = 65$) were also quite common.

[27] Note: We have 1080 voting decisions of public good players. However, in 291 cases, the sum of proposed punishment points from both third parties was the same. Therefore, we excluded these observations from

Table 3.
Success of third parties in Fourth Party and Competition

Variable	(1) Fourth Party	(2) Competition
Punishment of norm deviators	0.341	-0.189
	(0.383)	(0.203)
Average punishment within group	0.495***	-0.423***
	(0.128)	(0.108)
One player punished	0.426*	0.883***
	(0.249)	(0.193)
Constant	0.337	0.164
	(0.318)	(0.142)
Observations	360	720
Wald-χ^2	27.32***	39.76***

Notes: Logit regressions with subject random effects. 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) ($yes = 1/no = 0$). Robust standard errors, clustered at the subject level, in parentheses. Punishment of norm deviators: third party punished all public good players who contributed less than group average ($yes = 1/no = 0$). *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

4.3 Contributions and earnings

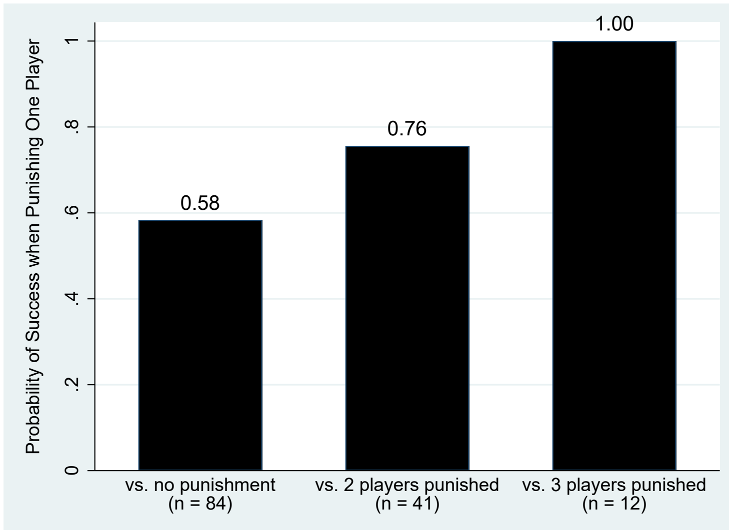
We now turn to the contribution decisions of the public good players. It was already clear from Figure 2 that the level of contributions depended on the treatment. For a more detailed analysis, we ran three random effects tobit regressions for total contributions, which we report in columns (1)-(3) in Table 5. The regressions confirm that contributions were significantly higher in Fourth Party and Competition than in the benchmark Baseline treatment. Wald-tests show that the difference between *Fourth Party* and *Competition* is statistically significant at the 10% level in (1) and (2), but not significant in (3) (see last row of Table 5). In addition, regression (2) shows that contributions went somewhat down in later periods. Regression (3) reveals that a higher average contribution of the other group members in the previous period had a significant positive effect on contributions. Overall, these results suggest that our sample contains conditional cooperators who base their contribution decision (in part) on the observed behavior of others.

Next, we analyze contribution behavior at the group level. During the 20 periods of the experiment, there were 240 groups in Baseline, 360 groups in Fourth Party and 360 groups in Competition. Calculating the standard deviation of contributions for each group and then calculating the mean of this variable across all groups and periods yields

the regression presented in Table 4.

[28]Additionally, period has a significant positive effect on whether public good players vote for the candidate who proposes more punishment.

Fig. 6. Probability of success when punishing one player in Competition ("matchups")



6.46 for Baseline, 4.87 for Fourth Party, and 4.41 for Competition. Wilcoxon rank sum tests confirm that the differences in mean standard deviation between Baseline and Competition ($p < 0.01$) and between Fourth Party and Competition ($p < 0.05$) are significant.[29]

Finally, we analyze the earnings for public good players: Which system is most profitable from their point of view? We do not include the earnings of third parties in this analysis because we consider them "outsiders" in our experiment in the sense that they do not benefit from high contributions to the public good. Therefore, we focus only on the payoffs of the public good players who constitute our "society" in the laboratory. Figure 7 shows the average earnings of public good players by treatment, revealing them to be highest in Competition, followed by Fourth Party and Baseline. Column (4) of Table 5 shows the results of a random effects regression in which the period-specific profits of the public good actors are the dependent variable. We again use Baseline as the benchmark. The results of this regression and a Wald test between *Fourth party* and *Competition* show that the returns in Competition were significantly higher than in Baseline and Fourth party.[30]

Result 4. *Contributions were higher in* Fourth Party *than in* Baseline*, in line with Hypothesis 1. We found them to be also higher in* Competition *compared to* Baseline*.*

[29] Table C3 in Appendix C shows the mean standard deviation of group-level contributions over the 20 periods after treatment.

[30] When clustering standard errors by session (see Table B3, Appendix B), the results support our findings that earnings were significantly higher in Competition than in Baseline. However, the difference between *Fourth Party* and *Competition* turns out not to be significant. Wilcoxon rank sum tests with session averages yield similar results: Baseline vs. Fourth Party ($p = 0.248$), Baseline vs. Competition ($p = 0.083$), Fourth Party vs. Competition ($p = 0.275$).

Table 4.
Voting behavior of public good players in COMPETITION

Variable	(1) Vote for more punishment
Contribution	0.248***
	(0.024)
Period	0.056***
	(0.015)
Constant	-3.603***
	(0.322)
Observations	789
Wald-χ^2	110.850

Notes: Logit regression with subject random effect. Dependent variable: Vote of public good player for the candidate who proposed more total punishment points for the group ($yes = 1/no = 0$). Robust standard errors, clustered at the subject level, in parentheses. *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

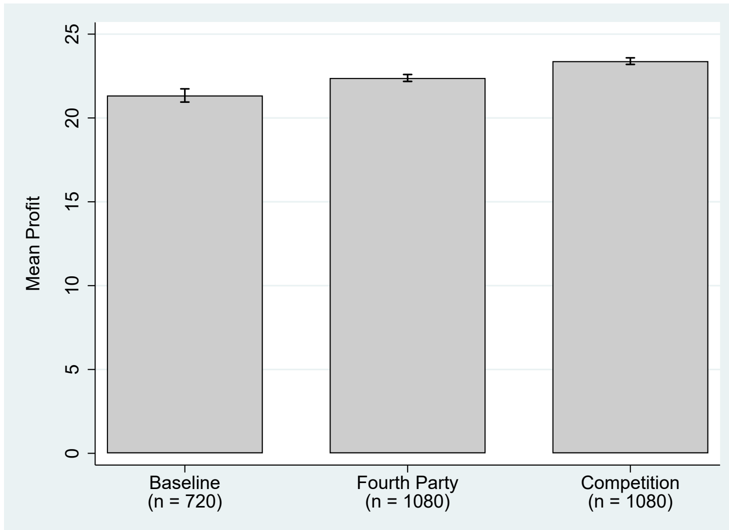
Moreover, the COMPETITION *treatment generated the highest earnings for public good players.*

5 Concluding remarks

Our study demonstrates significant differences in how third parties carry out their role depending on the presence and nature of additional layers of incentives. In our laboratory experiment, the introduction of an independent fourth party performed best if the objective is to reduce free-riding among public good players. However, this came at a high cost of punishment. On the other hand, electoral competition between two candidates for norm enforcement authority resulted in the greatest earnings for public good players among the three options we considered. In our experiment with three-player groups involved in a PGG, this institutional set-up endogenously led to focused sanctions against the least contributor.

Experimental studies where players could directly choose between automatic punishment of the least contributor and peer-to-peer punishment found the former to be popular and sufficient to maintain cooperation (see Andreoni and Gee 2012; Kamijo et al. 2014; see also Nicklisch et al. 2016 on endogenous choice of punishment regime in a PGG). We caution, however, that, as our theoretical model indicates, the effect of COMPETITION on outcomes can be highly sensitive to the moral preferences in the group of public good players, in particular to the moral standards of the median individual. Our findings also complement results from the second-party punishment literature, which show that voting

Fig. 7. Earnings of public good players by treatment



Notes: Error bars represent 95% confidence intervals for the mean.

allows subjects to agree on more efficient punishment schemes (see, e.g., Ertan et al. 2009; Putterman et al. 2011; Markussen et al. 2014).

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 potential 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 for now concerns the endogenous selection of an incentive environment for third parties. Our findings here suggest that this choice may well depend on various factors such as preferences for efficiency, but also on the beliefs about others' moral standards.

Table 5.
Contributions and earnings of public good players

Variable	(1)	(2) Contributions	(3)	(4) Earnings
Fourth Party	7.385***	7.397***	6.469***	1.044*
	(1.669)	(1.668)	(1.673)	(0.578)
Competition	4.881***	4.895***	4.135**	2.047***
	(1.666)	(1.666)	(1.669)	(0.590)
Period		-0.084***	-0.087***	-0.021
		(0.016)	(0.017)	(0.016)
Average contributions of others $(t-1)$			0.231***	
			(0.025)	
Constant	4.643***	5.509***	4.020***	21.558***
	(1.301)	(1.311)	(1.313)	(0.578)
Observations	2880	2880	2736	2880
Wald-χ^2	19.674***	46.401***	142.387***	21.846***
Fourth Party = *Competition*	$p = 0.0894$	$p = 0.0896$	$p = 0.1141$	$p = 0.0001$

Notes: Tobit regressions with subject random effects 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), we use robust standard errors, which are clustered at the subject level. $*p \leq 0.10$. $**p \leq 0.05$. $***p \leq 0.01$.

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.

Baldi, G. and C. LaFrance (2013). Lessons from the United States sheriff on the electoral selection of police commissioners in England and Wales. *Policing: A Journal of Policy and Practice 7*(2), 148–157.

Bandyopadhyay, S., and B. McCannon (2014). The effect of the election of prosecutors on criminal trials. *Public Choice 161*, 141–156.

Besley, T., and S. Coate (2003). Elected versus appointed regulators: Theory and evidence. *Journal of the European Economic Association 1*(5), 1176–1206.

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

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.

Cubitt, R., M. Drouvelis, S. Gächter, and R. Kabalin (2011). Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics 95 95*(3-4), 253–264.

d'Adda, G., M. Dufwenberg, F. Passarelli, and G. Tabellini (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior 124*, 288–304.

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

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.

Feinberg, M., R. Willer, J. Stellar, and D. Keltner (2012). The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology 102*(5), 1015–1030.

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.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences 35*, 1–59.

Hainmueller, J., and D. Hangartner (2019). Does Direct Democracy Hurt Immigrant Minorities? Evidence from Naturalization Decisions in Switzerland. *American Journal of Political Science 63*(3), 530–547.

Huber, G., and S. Gordon (2004). Accountability and coercion: Is justice blind when it runs for office? *American Journal of Political Science 48*(2), 247–263.

Jordan, J., K. McAuliffe, and D. Rand (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics 19*, 741–763.

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. (2021). Teams do inflict costly third-party punishment as individuals do: Experimental evidence. *Games 12*(1), 22.

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.

Kunkler, A. (2020). Elected or appointed? King County weighs sheriff options. https://www.seattleweekly.com/news/elected-or-appointed-king-county-weighs-sheriff-options/.

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

Lim, C. (2013). Preferences and Incentives of Appointed and Elected Public Officials: Evidence from State Trial Court Judges. *American Economic Review 103*(4), 1360–1397.

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.

Partridge, M., and T. Sass (2011). The productivity of elected and appointed officials: the case of school superintendents. *Public Choice 149*, 133–149.

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.

Wissel, P. (2020). King County Council takes steps to appoint next sheriff. https://www.knkx.org/post/king-county-council-takes-steps-appoint-next-sheriff.

Whalley, A. (2013). Elected versus Appointed Policy Makers: Evidence from City Treasurers. *The Journal of Law & Economics 56*(1), 39–81.

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

Zoorob, M. (2019). There's (rarely) a new sheriff in town: The incumbency advantage for county sheriffs. *Available at https://ssrn.com/abstract=3485700*.

A Proofs

Proof of Proposition 1

The third party's punishment decisions minimize her loss (2). The interior solution of this minimization problem must satisfy the first-order conditions

$$\kappa - \frac{1}{\theta_T}(N^{\text{Base}} - c_i - p_i) = 0 \qquad \text{for each } i\,.$$

It is straightforward to see that the second-order condition holds. The optimal punishment imposed on an agent i who contributed $c_i \geq 0$ is thus as stated in (7).

Since $\mathbf{E}(\theta_T) = \frac{1}{2}$, a public good player can in expectation avoid punishment by contributing $c_i^0 = N^{\text{Base}} - \frac{\kappa}{2}$. Anticipating the third party's punishment behavior, a public good player chooses c_i to minimize

$$\mathcal{L}_i = c_i + \lambda \max\left(0, N^{\text{Base}} - c_i - \frac{\kappa}{2}\right) + \frac{(N^{\text{Base}} - c_i)^2}{2\theta_i} + \sum_{\substack{j \neq i \\ j \in S}} \frac{(N^{\text{Base}} - c_j - p_j)^2}{2\theta_i}\,.$$

This can be transformed into the standard minimization problem $\min L(c_i, z)$ where

$$L(c_i, z) = c_i + \lambda z + \frac{(N^{\text{Base}} - c_i)^2}{2\theta_i} + \sum_{\substack{j \neq i \\ j \in S}} \frac{(N^{\text{Base}} - c_j - p_j)^2}{2\theta_i}\,,$$

subject to inequality constraints

$$z \geq 0$$

$$\text{and} \qquad z - N^{\text{Base}} + c_i + \frac{\kappa}{2} \geq 0\,.$$

The solutions for this problem imply that $c_i = N^B - \theta_i$, or $c_i = N^B - \theta_i(1-\lambda)$ or $c_i = N^B - \frac{\kappa}{2}$. Comparing the optimal values of the objective function (3) over the domain of θ_i, we arrive at expression (6).

Using equilibrium contributions (6), the expected value with respect to the distribution G of types is

$$\mathbf{E}^{\text{Base}}(c) = N^{\text{Base}} - \frac{\kappa^2}{8}\left(\frac{1}{1-\lambda} - 1\right) - \frac{1-\lambda}{2}\,. \tag{A.1}$$

In equilibrium, condition (1) has to hold. The equilibrium norm stated in (8) follows from inserting (A.1) into (1) and solving. $\square$

Proof of Proposition 2

As shown in Proposition 1, a third party's most preferred punishment in the absence of a fourth party is $\max\{0, N^I - c_i - \theta_T \kappa\}$. Third parties with type $\theta_T \geq \sqrt{\delta}/\kappa$ prefer less punishment compared to what is necessary to avoid being sanctioned by F. For these types, minimizing loss (4) thus calls for punishing all public good players who contributed strictly less than $N - \sqrt{\delta}$ by $N^F - c_i - \sqrt{\delta}$, i.e., the amount of punishment that makes F indifferent between sanctioning and not sanctioning T.

Third party types with $\theta_T < \sqrt{\delta}/\kappa$ prefer punishment $N^{FP} - c_i - \theta_T \kappa$, which satisfies F. Public good players in equilibrium anticipate third party behavior., they expect to be punished

$$p_i = \begin{cases} \max\{0, N^F - c_i - \sqrt{\delta}\} & \text{if } \theta_T \geq \frac{\sqrt{\delta}}{\kappa} \\ \max\{0, N^{FP} - c_i - \theta_T \kappa\} & \text{if } \theta_T < \frac{\sqrt{\delta}}{\kappa} \,. \end{cases}$$

Using that θ_T is drawn from a uniform distribution on the unit interval, the expected punishment of public good player i is

$$\mathbf{E}(p_i|\mathbf{c}) = \max\{0, N^{FP} - c_i - \sqrt{\delta} + \frac{\delta}{2\kappa}\} \,.$$

A public good player chooses c_i to minimize

$$\mathcal{L}_i = c_i + \lambda \max\left(0, N^{\mathrm{FP}} - c_i - \sqrt{\delta} + \frac{\delta}{2\kappa}\right) + \frac{(N^{\mathrm{FP}} - c_i)^2}{2\theta_i} + \sum_{j \neq i} \frac{(N^{\mathrm{FP}} - c_j - p_j)^2}{2\theta_i} .$$

This can be transformed into the standard minimization problem $\min L(c_i, z)$ where

$$L(c_i, z) = c_i + \lambda z + \frac{(N^{\mathrm{FP}} - c_i)^2}{2\theta_i} + \sum_{j \neq i} \frac{(N^{\mathrm{FP}} - c_j - p_j)^2}{2\theta_i} \,,$$

subject to inequality constraints

$$z \geq 0$$

$$\text{and} \qquad z - N^{\mathrm{FP}} + c_i + \sqrt{\delta} - \frac{\delta}{2\kappa} \geq 0 \,.$$

The solutions for this problem imply that $c_i = N^{\mathrm{FP}} - \theta_i$, or $c_i = N^{\mathrm{FP}} - \theta_i(1-\lambda)$ or $c_i = N - \sqrt{\delta} + \frac{\delta}{2\kappa}$. Comparing the optimal values of the objective function (5) over the domain of θ_i, we arrive at expression (9).

Using equilibrium contributions (9), the expected value with respect to the distribution G of types is

$$\mathbf{E}^{\mathrm{FP}}(c) = N^{\mathrm{FP}} - \frac{\delta\lambda + (1-\lambda)^2}{2(1-\lambda)} + \frac{\lambda}{(1-\lambda)} \left(\frac{\sqrt{\delta}\delta}{2\kappa} - \frac{\delta^2}{8\kappa^2} \right) \,. \tag{A.2}$$

In equilibrium, condition (1) has to hold. The equilibrium norm stated in (11) follows from inserting (A.2) into (1) and solving. It follows from comparing (11) and (8) that $\sqrt{\delta} < \kappa/2$ is necessary and sufficient to ensure that $N^{\text{FP}} > N^{\text{Base}}$ for all $\beta > 0$ and $\lambda > 0$. $\square$

Proof of Proposition 3

We solve by backward induction. Let $\mathcal{L}_i^A$ and $\mathcal{L}_i^B$ denote the losses that public good player i incurs if, respectively, candidate A's punishment proposal $\mathbf{p}^A = (p_1^A, \ldots, p_l^A)$ or candidate B's proposal $\mathbf{p}^B = (p_1^B, \ldots, p_l^B)$ is implemented. Clearly, i will cast his vote in favor of candidate A (B) if $\mathcal{L}_i^A < (>) \mathcal{L}_i^B$. In case that $\mathcal{L}_i^A = \mathcal{L}_i^B$, voter i is indifferent and decides by the toss of a coin.

The loss function (3) implies that, for any own contribution c_i, i's loss is at a minimum when $p_i = 0$ and $p_j = N^{\text{Comp}} - c_j$ for all $j \neq i$. In the equilibrium of majority competition, both office-motivated candidates thus propose to punish all public good players who contribute strictly less than the median amount $\tilde{c}_M$. Public good players who contribute at least $\tilde{c}_M$ face zero punishment. It follows from the loss mimization problem that the non-punished players' optimal contribution is $c_i = N^{\text{Comp}} - \theta_i$. The expected median contribution thus is $c_M = N^{\text{Comp}} - \theta_M$, i.e., the loss-minimizing contribution of a public good player with median type θ_M.

An office-seeking third party candidate has to disregard his own preferences (his θ-type) and choose punishment in line with the preferences of group members who contributed at least the median amount. We thus arrive at expression (13).

Public good players with type $\theta_i > \theta_M$ expect to face punishment $p_i = \max\{0, N^{\text{Comp}} - c_i\}$. Minimizing loss in face of non-zero punishment, i's optimal contribution is $c_i = N^{\text{Comp}} - \theta_i(1-\lambda)$. Clearly, a public good player of type $\theta_i > \theta_M$ prefers to avoid punishment iff

$$N^{\text{Comp}} - \theta_M \leq N^{\text{Comp}} - \theta_i(1-\lambda) \quad \text{or, equivalently,}$$
$$\theta_i \leq \frac{\theta_M}{1-\lambda} \stackrel{\text{def}}{=} \tilde{\theta}\,.$$

This gives us the optimal contributions stated in part (i) of the proposition.

Next, we show part (iii). Using equilibrium contributions (12), the expected value with respect to the distribution G of types is

$$\mathbf{E}^{\text{Comp}}(c) = N^{\text{Comp}} - \frac{\theta_M^2}{2}\left(\frac{1}{1-\lambda} - 1\right) - \frac{1-\lambda}{2}\,. \tag{A.3}$$

In equilibrium, condition (1) has to hold. The equilibrium norm stated in (14) follows from inserting (A.3) into (1) and solving.

To show part (iv), compare the marginal effect on i's loss (3) from being punished one unit more to the marginal effect of other players being punished one unit more. Being punished one unit more increases i's loss by λ, whereas each unit of punishment imposed on another group member $j \neq i$ reduces i's loss by

$$\frac{N^{\textsc{Comp}} - c_j - p_j}{\theta_i}\,.$$

This shows that condition (15) is sufficient to conclude that i derives a greater disutility from being punished himself than he derives utility from other norm-violators being punished. If the condition is satisfied and the candidates' proposals differ, then i will prefer the candidate who proposes to punish him less. $\square$

B Regressions with session clustered standard errors

Table B1.
Estimates for individually assigned punishment (session clustered standard errors)

Variable	(1) Punishment decision	(2) Punishment decision	(3) Punishment level	(4) Punishment level
Fourth Party	1.394***	1.266**	0.868	-0.038
	(0.482)	(0.601)	(0.555)	(0.685)
Competition	0.449	0.311	0.865	0.638
	(0.409)	(0.369)	(0.598)	(0.785)
Deviation	-0.258***	-0.146***	-0.333***	-0.144**
	(0.027)	(0.020)	(0.058)	(0.057)
Fourth Party x Deviation		-0.285***		-0.299***
		(0.072)		(0.056)
Competition x Deviation		-0.092***		-0.113
		(0.020)		(0.069)
SD of group contributions		0.047*		0.104***
		(0.028)		(0.034)
Constant	-1.363***	-1.377***	1.354**	1.394**
	(0.416)	(0.390)	(0.661)	(0.691)
Observations	3960	3960	1397	1397
Wald-χ^2	97.087***	6141.259***	42.643***	25230.036***
Fourth Party = *Competition*	$p = 0.0003$		$p = 0.9939$	
Fourth Party x Deviation = *Competition* x Deviation		$p = 0.0066$		$p = 0.0000$

Notes: (1) and (2) are probit regressions with subject random effects where the dependent variable takes a value of 1 if the B-player punished an A-player and 0 otherwise. (3) and (4) are 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 session level in all 4 regressions, in parentheses. Deviation: Individual contribution minus average contribution within the group. SD: Standard deviation. *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

Table B2.
Number of punished players in a group: regressions (session clustered standard errors)

Variable	(1) no punishment	(2) one player punished	(3) >one player punished
Fourth Party	-3.661***	1.105***	1.901***
	(0.759)	(0.237)	(0.504)
Competition	-1.298**	1.503***	-0.042
	(0.553)	(0.115)	(0.378)
SD of group contributions	-0.268***	0.095**	0.121***
	(0.064)	(0.042)	(0.022)
Period	-0.022	0.020	-0.008
	(0.033)	(0.022)	(0.018)
Constant	2.210***	-2.767***	-2.061***
	(0.784)	(0.444)	(0.313)
Observations	1320	1320	1320
Wald-χ^2	47.278***	244.962***	78.363***
Fourth Party = *Competition*	$p = 0.0000$	$p = 0.0914$	$p = 0.0000$

Notes: Logit regressions with subject random effects. 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 session level, in parentheses. SD: Standard deviation. *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

Table B3.
Earnings of public good players (session clustered standard errors)

Variable	(1) Earnings
Fourth Party	1.044
	(0.964)
Competition	2.047**
	(0.881)
Period	-0.021
	(0.032)
Constant	21.558***
	(0.860)
Observations	2880
Wald-χ^2	6.439*
Fourth Party = *Competition*	$p = 0.1289$

Notes: Random effects regression. Dependent variable: Earnings. Standard errors, clustered at the session level, in parentheses. *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

C Additional figures and regressions

Table C1.
Assigned punishment (group level)

	(1)	(2)	(3)	(4)
Variable	Punishment decision		Punishment level	
Fourth Party	2.325***	2.327***	0.703	0.703
	(0.588)	(0.644)	(0.545)	(0.606)
Competition	0.872*	0.925*	0.410	0.382
	(0.489)	(0.539)	(0.543)	(0.595)
SD of group contributions	0.148***	0.161***	0.136***	0.161***
	(0.024)	(0.028)	(0.022)	(0.023)
Sum of group contributions	-0.014**	-0.011**	-0.015***	-0.017***
	(0.006)	(0.005)	(0.004)	(0.004)
SD of group contributions ($t-1$)		-0.011		0.015
		(0.019)		(0.011)
Sum of group contributions ($t-1$)		0.000		0.002
		(0.006)		(0.004)
Period		0.012		0.019*
		(0.011)		(0.011)
WTP^a		0.842		-0.128
		(0.583)		(0.631)
WTP^d		0.131		0.031
		(0.415)		(0.465)
Constant	-0.844*	-1.363**	1.046*	0.665
	(0.479)	(0.686)	(0.536)	(0.751)
Observations	1320	1254	864	814
Wald-χ^2	55.239	55.104	60.867	79.737
Fourth Party = *Competition*	$p = 0.0000$	$p = 0.0001$	$p = 0.3488$	$p = 0.3151$

Notes: (1) and (2): Probit regressions with group random effect 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 group 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. C1. Number of punished players for third parties in FOURTH PARTY by assessment (fractions)

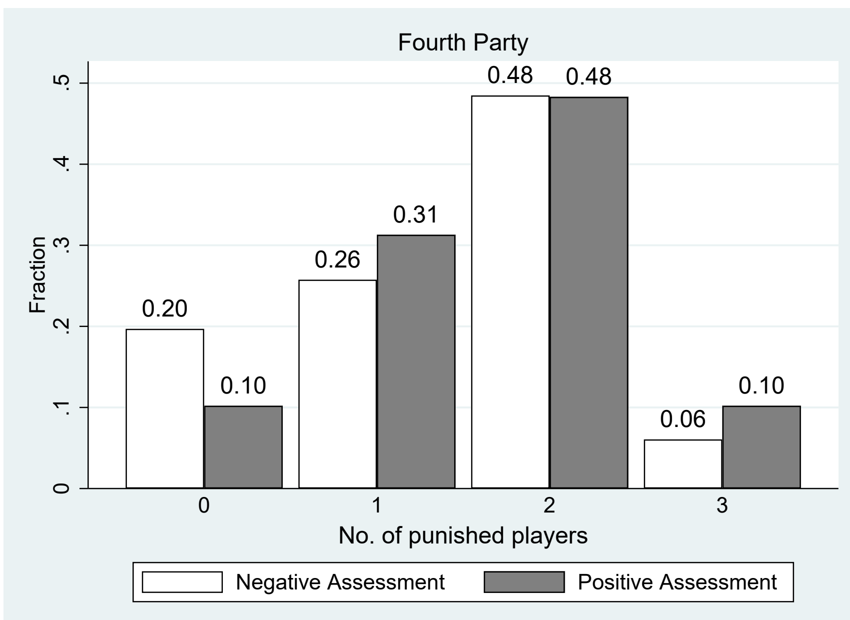


Table C2.
Success of third parties in Fourth Party and Competition (2)

Variable	(1) Fourth Party	(2) Fourth Party	(3) Competition	(4) Competition
Punishment of norm deviators	0.341		-0.293	
	(0.383)		(0.211)	
Average punishment within group	0.495***	0.586***	-0.409***	-0.485***
	(0.128)	(0.158)	(0.109)	(0.100)
One player punished (lowest contr.)	0.426*	0.485*	1.003***	0.893***
	(0.249)	(0.273)	(0.208)	(0.172)
Deviation to punishment profile		-0.007		0.002
		(0.014)		(0.013)
Constant	0.337	0.541**	0.181	0.107
	(0.318)	(0.250)	(0.136)	(0.137)
Observations	360	360	720	720
Wald-χ^2	27.318	19.143	42.938	46.200

Notes: Logit regressions with subject random effects. 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 players with contributions over 8; 1 punishment point for players with contributions of 8; 2 punishment points for players with contributions of 4; 3 punishment points for players with contributions of 0. *$p \leq 0.10$. **$p \leq 0.05$. ***$p \leq 0.01$.

Table C3.
Mean of standard deviation of contributions per group by period and treatment

Period no.	1	2	3	4	5	6	7	8	9	10
Baseline	6.72	8.32	6.29	6.88	6.04	7.71	7.45	6.80	5.73	7.10
Fourth Party	6.42	5.56	5.70	5.13	5.17	5.16	4.20	5.51	4.82	4.00
Competition	6.04	4.37	4.23	5.62	5.03	4.78	4.75	4.58	4.85	3.58
Period no.	11	12	13	14	15	16	17	18	19	20
Baseline	5.27	6.71	5.98	7.00	5.76	6.71	5.75	5.70	5.00	6.30
Fourth Party	5.16	4.45	5.16	5.01	4.90	4.18	4.15	4.75	4.03	4.01
Competition	4.31	4.20	4.03	4.10	4.35	3.60	3.40	3.90	4.42	4.01

D Instructions (English translations)

Welcome to the experiment and thank you for your participation.[31]

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.

D.1 Part 1 (all treatments)

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 D1.

Example: The left option in the second row in Table D1 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.)

[31] The original instructions were in German. Part 1 was identical in all treatments. The instructions for Part 2 of the treatments Baseline, Fourth Party and Competition are reported in Section D.2, D.3 and D.4, respectively.

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 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.

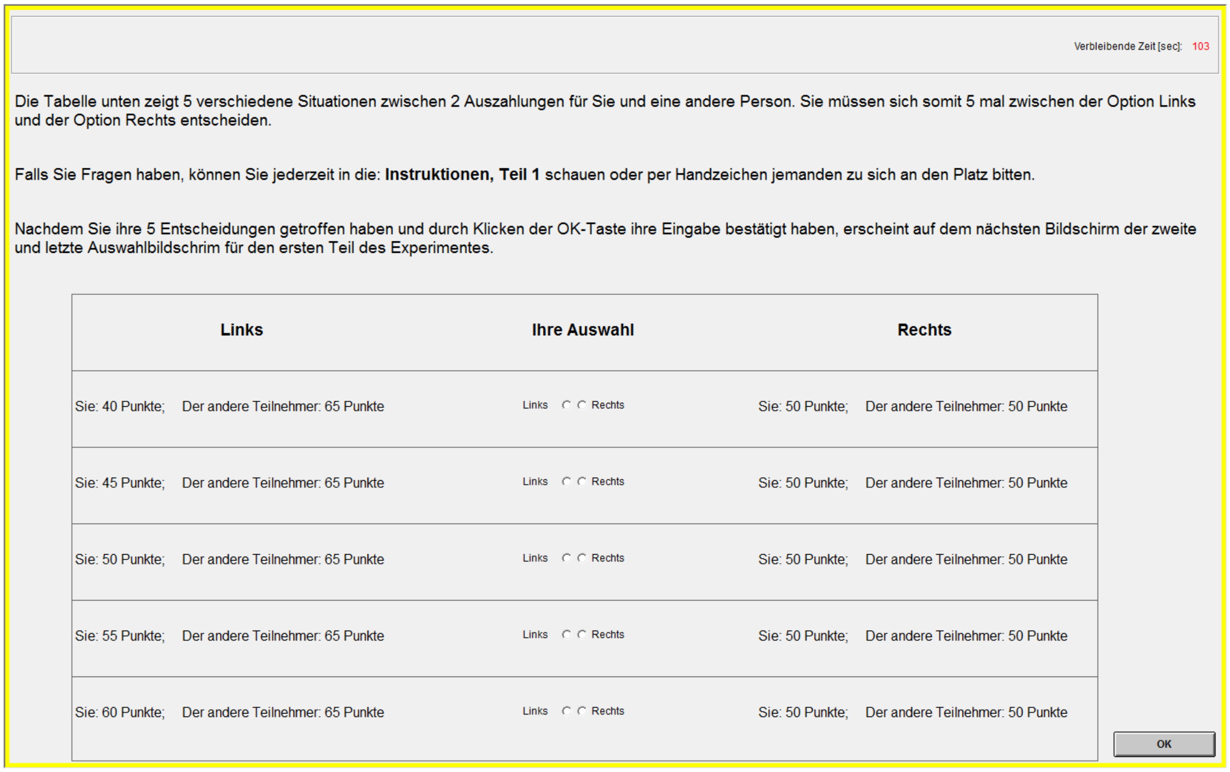
D.2 Part 2 (Baseline)

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

Fig. D1. Decision screen in Part 1



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 D2 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 D2 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 D2**.

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: $20 - 8 + 0.5 * (8 + 16) = 24$). 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 B-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.

D.3 Part 2 (Fourth Party)

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*, *B-player* or *C-player*). In total, there are 18 players of type A, 6 players of type B and 6 players of type C. 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 and one C-player are 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 their 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.

After the B-players have made their decision, the C-players are given information about the individual contributions of the players A1, A2 and A3, as well as the assigned punishment points of the B-player to A1, A2 and A3 in their group. Then, each C-player has to decide whether he considers the decision of the B-player in his group appropriate or not appropriate.

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 D2 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 D2 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 D2**.

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: $20-8+0.5*(8+16)=24$). 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 B-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. There are two possible cases:

If the decision of the B-player has been evaluated as **appropriate** by the C-player in his group, then the payoff of the B-player is calculated as follows:

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

If the decision of the B-player has been evaluated as **not appropriate** by the C-player in his group, then the payoff of the B-player is 5 points.

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, and the decision of the B-player has been evaluated as appropriate by the C-player, then the payoff of the B-player in this round would equal 20 points (calculation: $30 - 2 - 3 - 5 = 20$).

Payoff of C-players

As a C-player, your payoff equals 15 euros (= 500 points) for the entire Part 2, irrespective of your choices made in Part 2.

Final payoff in Part 2

After all A-, B- and C-players have made their decision in each round, the payoffs of A- and B-players for each round are calculated. At the end of each round, the A- and B-players receive information on how many points they earned. The final payoff of A- and B-players in Part 2 is calculated as the sum of the payoffs of all 20 rounds. The C-players get a payoff of 15 euros (= 500 points). At the end of Part 2, each participant receives information on his total payoff.

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.

D.4 Part 2 (Competition)

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 12 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, two B-players are randomly assigned to each of the six groups (Players B1 and B2). 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 their 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.

After the B-players have made their decision, the A-players are informed about the assigned punishment points of the two B-players in their group. Then, each A-player has to decide whether he prefers the decision of B1 or B2 in his group. Only the decision that is preferred by the majority of the A-players will be implemented in a group. A decision has to be preferred by at least

2 A-players of the group in order to be implemented and be relevant for payoff.

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 relevant B-player of the group. Each punishment point by the B-player who is preferred by the majority of the A-players 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 preferred by the majority * 2*

You cannot receive a negative payoff. If the formula yields a negative amount, your payoff is 0 points. Table D2 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 D2 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 D2**.

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: $20-8+0.5*(8+16)=24$). The payoff of the A-player in the current period equals 24 points minus the assigned punishment points of the B-player who is preferred by the majority of A-players 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 who is preferred by the majority of A-players in his group assigned 2 punishment points, the payoff of the A-player in this round would equal 30 points ($34 - 2 * 2 = 30$).

Payoff of B-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. There are two possible cases:

If the decision of the B-player is **preferred** by the majority of A-players in his group, then the payoff of the B-player is calculated as follows:

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

If the decision of the B-player is **not preferred** by the majority of A-players in his group, then the payoff of the B-player is 5 points.

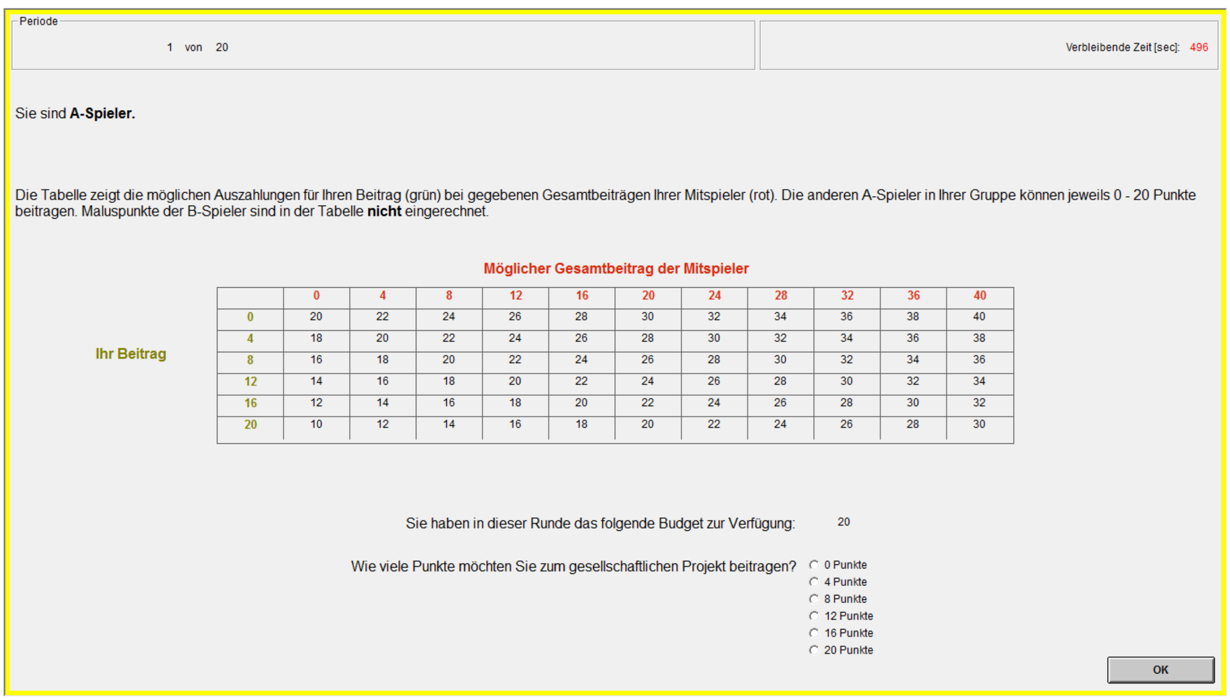
Example: If the B-player assigned 2 punishment points to player A1, 2 punishment points to player A2 and 5 punishment points to player A3, and the decision of the B-player is preferred by 2 out of the 3 A-players in his group (i.e., the majority of A-players), then the payoff of the B-player in this 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.

Fig. D2. Decision screen in Part 2


Periode

1 von 20

Verbleibende Zeit [sec]: 496

Sie sind **A-Spieler.**

Die Tabelle zeigt die möglichen Auszahlungen für Ihren Beitrag (grün) bei gegebenen Gesamtbeiträgen Ihrer Mitspieler (rot). Die anderen A-Spieler in Ihrer Gruppe können jeweils 0 - 20 Punkte beitragen. Maluspunkte der B-Spieler sind in der Tabelle **nicht** eingerechnet.

Möglicher Gesamtbeitrag der Mitspieler

Ihr Beitrag

	0	4	8	12	16	20	24	28	32	36	40
0	20	22	24	26	28	30	32	34	36	38	40
4	18	20	22	24	26	28	30	32	34	36	38
8	16	18	20	22	24	26	28	30	32	34	36
12	14	16	18	20	22	24	26	28	30	32	34
16	12	14	16	18	20	22	24	26	28	30	32
20	10	12	14	16	18	20	22	24	26	28	30

Sie haben in dieser Runde das folgende Budget zur Verfügung: 20

Wie viele Punkte möchten Sie zum gesellschaftlichen Projekt beitragen?

- 0 Punkte
- 4 Punkte
- 8 Punkte
- 12 Punkte
- 16 Punkte
- 20 Punkte

OK

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