


ORIGINAL ARTICLE

# Cooperation through collective punishment and participation

Dominik Duell<sup>1</sup> , Friederike Mengel<sup>2</sup>, Erik Mohlin<sup>3,4</sup> and Simon Weidenholzer<sup>2</sup>

<sup>1</sup>Department of Political Science, University of Innsbruck, Innsbruck, Austria, <sup>2</sup>Department of Economics, University of Essex, Colchester, UK, <sup>3</sup>Department of Economics, Lund University, Lund, Sweden and <sup>4</sup>The Institute for Futures Studies, Hölländargatan 13, 111 36 Stockholm, Sweden

Corresponding author: Dominik Duell; Email: [dominik.duell@uibk.ac.at](mailto:dominik.duell@uibk.ac.at)

(Received 6 January 2023; revised 26 May 2023; accepted 11 July 2023; first published online 12 December 2023)

## Abstract

We experimentally explore the role of institutions imposing collective sanctions in sustaining cooperation. In our experiment, players only observe noisy signals about individual contributions in finitely repeated public goods game with imperfect monitoring, while total output is perfectly observed as it is often the case in collective action problems in society. We consider sanctioning mechanism that allows agents to commit to collective punishment in case the level of cooperation among members of society falls short of a target. We find that cooperation is higher with collective punishment compared to both no punishment or punishment targeting individuals. Importantly, our results indicate that it is the combination of making a commitment to be punished and the collective nature of punishment which induces cooperation. Our findings show that punishing a group collectively for misbehavior of some of its members induces cooperation when individuals participate in setting up the sanctioning institution. The study contributes to the literature on institutional legitimacy and how to ensure good government performance when dealing with collective action problems, and, by considering commitment, improves enforcement methods criticized for their detrimental effects on some societal groups.

**Keywords:** collective action; collective sanctions; imperfect; participation; public goods game

## 1. Introduction

Sanctioning a group for the misconduct of some of its members is a common feature of how societies are governed. Particularly in situations where individual cooperation cannot be easily detected, collective punishment is an often employed enforcement regime. Historical examples include explicit punishment in military and educational institutions, such as the ancient Roman practice of decimation where one in ten soldiers of a military unit was randomly chosen to be executed. More contemporaneously, fencing off a park, restricting access to a private road, or imposing curfews during a pandemic as response to a small number of trespassers are frequently encountered policies. While the ability to sanction has been shown to increase cooperation among members of many types of groups (Ostrom *et al.*, 1992), collective sanctioning is deemed less effective because it is usually seen as unfair and crowds-out intrinsic motivation to comply. Collective punishment is particularly contested and often even said to be counterproductive when carried out as racial profiling in policing (Gelman *et al.*, 2007), indiscriminate retaliation for terror attacks (Bueno de Mesquita and Dickson, 2007), or economic sanctions imposed on rogue regimes that hurt the civilian population (Gordon, 1999; Allen and Lektzian, 2013). We know, however, that the origins and kind of authority to carry out punishment matters greatly (Dickson *et al.*, 2015, 2009), that sanctioning institutions established

© The Author(s), 2023. Published by Cambridge University Press on behalf of EPS Academic Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

by a community through elections is seen as effective in inducing high levels of cooperation (Miguel and Gugerty, 2005; Grossman and Baldassarri, 2012), in particular when democratically chosen authorities stand repeated elections (Castillo and Hamman, 2021). And, we know that individual actors are very willing to cooperate when the sanctioning institution is committed to credibly carry out enforcement—especially in the long-run—as stipulated by its rules (Baldwin, 2013, 2019), or when such institution arose endogenously (Kosfeld *et al.*, 2009; Markussen *et al.*, 2014).

This leads us to ask, particularly in situation where individual behavior is hard to observe, whether collective sanctioning can be effective in inducing successful collective action when it is paired with commitment to be punished as a group as well as with individuals' participation in the decision to establish such sanctioning institutions. Commitment to be punished constitutes self-selection of individuals into a group or institution that is governed by a sanctioning regime. Such institutions frequently exist: citizens in democratic regimes are subject to collective sanctions (e.g., restrictions to consuming certain goods as response to misbehavior of a few) but would have the option to exit (Warren, 2011) from this institutionalized pool punishment institution (Hilbe *et al.*, 2014), where individualized punishment is often explicitly forbidden (Balafoutas and Nikiforakis, 2012). Other examples of organizations featuring self-selection into collective sanctioning mechanism include labor unions engaged in collective bargaining where all of the union can be held liable for the breach of industrial action regulations by some of its members (Bruun and Johansson, 2014), credit unions where a few defaulting on their loans increases fees for all members (Wheelock and Wilson, 2013), or condominium associations, which force their members to build a fund for repairing damages incurred through the negligence of some (Van Der Merwe, 2015). Importantly, the threat to carry out the collective sanction is credibly ensured in any of these institutions, at least in developed societies, through legal provisions and contracts.

To answer whether collective sanctioning with commitment benefits collective action, we study finitely repeated public goods games in a laboratory experiment where individual actors receive noisy signals on the contributions of others but the sum of contributions is observed without noise. Our treatments then vary the punishment technology individual actors have at their disposal. We compare individuals' contribution and groups' welfare under the possibility for the individual actors to commit to an *ex-ante* known regime of collective punishment of the whole group to: simple collective punishment by a randomly chosen authority from among the group; peer-to-peer punishment committed to by each individual actor and targeting the individual for their behavior; non-committed peer-to-peer punishment; or, no punishment altogether. The laboratory allows to make all individual action, including non-cooperation, observable to the researcher—a situation that rarely occurs in the real world.

Imagine any group that aims to cooperate for achieving successful collective action, be it a democratic society, a labor union, a credit union, condominium association, or any group of individuals with a common objective; further, consider a situation where information about the contribution of each individual to the common objective is not perfect. The sanctioning technologies we instantiate in the laboratory, then, mimics incentives that exist in interactions in all such groups in the real world. What are these technologies and incentives? Individuals may punish their peers on their own accord. Such peer-to-peer punishment could be based in law (e.g., as a result of a civil court decision) or simply public shaming for not working toward the collective goal. A game-theoretic analysis will tell us that, in equilibrium, no punishment should ever be carried out and no contributions to the public goods should be made. At the point of decision whether to contribute to the group goal, group members simply cannot credibly commit to punishing a perpetrator *ex-post*. Empirically, we often observe such individual punishment, often even in excess and of those who actually contributed, leading to detrimental outcomes for the welfare of the group. Avoiding the empirically observed excessive punishment but also the theoretically suggested commitment problem, a solution for this might be if we ask individuals to commit to acting against their peers for not putting sufficient amount of effort into reaching common goals. Committing to such punishment, in equilibrium, leads only to increased

compliance if the signals about who is free-riding or in violation of what is being expected is precise. In many cases, as argued, our information about who is not cooperating is not precise. Which co-worker steals from the fridge in the common room, which neighbor is dumping trash beyond what is taken away by public services, or which resident is trespassing in a natural preserve is often not observed, only the consequences of those actions are. Once we lack information, peer-to-peer punishment, even if individuals committed to it ex-ante, will often fail to be carried out because we may not be able to identify the individual perpetrator. In these situations, if sanctions ought to be applied, punishing everyone is left as the only feasible option. The fridge in the common room is taken away, public services increase fees for the whole building, and the natural preserve is fenced off. Will such collective punishment increase cooperation? The theoretical expectation would be no, since contributions are predicted to stay at zero and no punishment is carried out. Again, empirically, we observe punishment of all, with similarly bad consequences for the welfare of the group as with peer-to-peer punishment but additionally putting a burden on the group and the institutions governing it because many contributors may be sanctioned as well. Collective committed punishment solves, once more, the commitment problem but will only avoid the problematic sanctioning of contributors if it induces high enough levels of contributions leading to no punishment.

Indeed, we find empirically that both contribution rates and welfare are higher in the presence of the possibility to commit to collective punishment than under either the absence of the possibility to punish or non-committed peer-to-peer punishment. Further, while welfare levels are constant in the presence of the possibility to commit to collective punishment, they are steadily decreasing in the other two cases. This points toward positive long-run welfare effects of being able to commit to collective punishment. The combination of the participation in the decision-making whether punishment occurs and social commitment to the punishment outperforms any other form of sanctioning or the absence of punishment altogether in sustaining cooperation. While contribution and payoff levels are in line with the no-punishment benchmark when no punishment is committed, they are significantly higher when punishment is committed.

Our study speaks to a broad set of questions: which institutions are seen as more legitimate, which institutions enable good government performance, and which institutions reduce inequalities in policy outcomes? The ability to participate in the decision-making process of institutions that govern individual lives increases system support (Finkel, 1987; Bowler and Donovan, 2002). An observation that has been experimentally tested with respect to what sanctioning institutions are seen as legitimate and which are successful in ensuring cooperation (Dickson *et al.*, 2009, 2015; Grossman and Baldassarri, 2012). We track whether individuals' repeated willingness to vote for setting-up collective sanctions, a measure of support of the punishment regime, sustains high levels of cooperation, providing a controlled test of the link between levels of system support and regime characteristics such as collectiveness and participation. A large literature sees participation as driver of particularly effective government (Fung and Wright, 2001; Agrawal, 2005) but concerns are raised whether involvement of too many stakeholders creates bad policy outcomes due to, for example, being elite dominated (Mansuri and Rao, 2004) or facing a trade-off between equity and efficiency (Hong and Cho, 2018). The institutions we create in the laboratory vary whether individuals are involved in setting them up, directly testing whether it is the participation in the decision how one is governed, i.e., is collective sanctioning implemented, and holding equal power over choosing the institution, that determines the quality of outcomes. We also provide new avenues, by considering commitment, to improve those methods of enforcing a policy that are criticized for their detrimental effect on some societal groups and are demonstrated to be welfare inefficient in general, such as any kind of profiling or zero-tolerance policies (for a review see Soss and Weaver, 2017). We deliver empirical evidence how commitment to collective sanctions sets incentives to monitor and control wrongdoers within one's group of peers as debated in international law in particular, as the questions of how to punish groups committing mass atrocities (Drumbl, 2004), and legal studies in general as the question whether delegation of

detering wrongdoing to the group creates better outcomes because it gives agency (Levinson, 2003).

We also add to the experimental literature in political economy, economics, and other behavioral science disciplines that examines factors that may lead to the emergence of high effort equilibria in coordination games that model collective action problems (Brandts and Cooper, 2006; Weber, 2006; Feri *et al.*, 2010; Henrich *et al.*, 2010; Riedl *et al.*, 2015). We demonstrate that collective punishment transforms the underlying public good game into a coordination game. We contribute to this literature by showing that high effort equilibria may also be reached when the coordination game is induced by the choice of players, rather than being exogenously given.

### **1.1 Cooperation under sanctioning with imperfect information**

Ensuring compliance or cooperation through sanctioning is generally seen as effective. It is well-documented in an experimental context that the ability to punish allows societies to move toward socially optimal levels of compliance and contributions (see, e.g., Ostrom *et al.*, 1992; Fehr and Gächter, 2000, 2002; Gintis *et al.*, 2005). Even a randomly assigned individual singled out to punish defectors is able to increase contribution and welfare levels (Boyd and Richerson, 1992; O’Gorman *et al.*, 2009). Despite costly punishment being at odds with selfish preferences, a considerable fraction of individuals is willing to sacrifice own payoffs to punish non-cooperators, a pattern of behavior also observed in many populations around the world (Henrich *et al.*, 2010).

A key prerequisite for punishment to be effective is the ability to identify non-contributors. If it is not clear who the free-riders are, it becomes impossible to accurately target them. Perhaps even worse, from time to time it will be the case that some contributors are falsely identified as non-contributors and therefore punished. Indeed, a decline in contributions and welfare under imperfect monitoring is well documented in the experimental literature (see e.g., Bornstein and Weisel, 2010; Grechenig *et al.*, 2010; Ambrus and Greiner, 2012). Such imperfect monitoring, more generally, is one of the main reasons why voters are not able to induce behavior among elected officials that benefits society (Powell and Powell Jr, 2000; Tavits, 2007) or for governments, as central authority, to make citizens to comply with the law or to contribute to public goods easily even under the threat of sanctioning. A government’s authority to sanction is eroded, in the eyes of citizens, and the incentive for non-compliance rises when imperfect monitoring leads to increased punishment of cooperators (Dickson *et al.*, 2009). Even when the punishment comes from one’s peers, sanctioning cooperators for their pro-social behavior demoralizes and renders punishment as a tool to induce cooperation useless (Herrmann *et al.*, 2008). Considerations along these lines already two centuries ago might have lead Blackstone (1966) to assert that “the law holds that it is better that ten guilty persons escape than that one innocent suffers.”

### **1.2 Cooperation under a centralized authority or with collective punishment**

In contexts where the state is empowered to sanction, such centralized authority is often said to be effective. Such effectiveness arises, for example, through the legitimacy of the central authority to carry out punishment (Baldassarri and Grossman, 2011) or by way of domestic backlash for the regime leader sanctioned by an international authority (McGillivray and Smith, 2000, 2006). Centralizing the sanctioning authority in governmental institutions, however, is often outdone by decentralized punishment exactly because misattribution is less frequent in the latter. Punishing effectively relies on the legitimacy of the sanctioning institution (Dickson *et al.*, 2009; Eckel *et al.*, 2010) undermined by punishing cooperators and not punishing culprits. Decentralization allows for locally informed punishment avoiding that the bluntness of tools available to a central authority is too clumsy (Ostrom, 1999). The positive effect of when punishment takes place at the local level even extends back to the centralized authority when it is legitimized internally, for example, through elections or appointment by the monitored communities (Miguel and Gugerty, 2005; Grossman and Baldassarri, 2012).

Independent of whether sanctioning is done by a centralized or decentralized authority, the target of punishment, the individual actor alone or the whole group, matters as well. Peer-to-peer punishment is effective but mostly only if done in small groups (Sigmund, 2007; Taylor, 1982) and it needs the expectation of future interactions for sanctioning to be a threat, an expectation that is mute in many large-scale societal interactions (Boyd *et al.*, 2010). Legal studies see collective punishment as a solution to this problem, and deem such sanctioning tools as justified as well as effective in ensuring compliance and cooperation (Levinson, 2003). Collective incentives are also an important part of salary schemes in the form of team incentives (Ledford *et al.*, 1995). There is also some empirical evidence that receiving a bonus lower than an expected amount decreases individuals' satisfaction (Ockenfels *et al.*, 2014) and that collective punishment is experienced as unfair (Chapkovski, 2021). Additionally, group outcomes, which are typically observed in a much less noisy way, anchor sanctions with collective punishment making the problem of misattribution more severe (Holmström, 1982). Despite these drawbacks, collective punishment may still play an important role by potentially enhancing welfare levels that is the overall benefits to all members of society from the plus in cooperation induced by sanctioning.

### 1.3 Collective punishment with commitment

Accepting for now that collective punishment potentially overcomes some of the shortcomings of punishment targeting the individual, even if it generates a different set of problems, which form of sanctioning then could ensure that collective punishment is seen as legitimate and therefore effective in triggering cooperation? From above, we take that the process by which an authority obtains the right to punish is important (Dickson *et al.*, 2009; Grossman and Baldassarri, 2012) as well as how this monitoring institution benefits from cooperation of others; individuals tend to show an indirect free-rider aversion, where they want to give less if the monitoring institution makes money contingently on the cooperation of others (Alventosa *et al.*, 2021).

Participation in the decision-making process whether, whom, and how to sanction has also been shown to improve cooperation, sometimes by helping the enforcement of the punishment decision as in Dickson *et al.* (2009), by sustaining compliance (Cox *et al.*, 2010), or by ensuring the survival of the sanctioning institution itself (Falleti and Riofrancos, 2018). Still, mixed evidence exists whether voting on the sanctioning authority could be an example of such beneficial participation. DeCaro *et al.* (2015) find that the ability to vote on rules improves cooperation but only if enforcement of compliance is certain, Chang *et al.* (2018) add that enforcement through an elected authority is often undermined by, for example, political inequality across individual actors, and Heap *et al.* (2020) find that including individuals in collective decisions through giving "voice" increases cooperation more than through participation in form of a vote. Baldwin (2013, 2019) even show that local decision-makers with centralized authority to punish, i.e., chiefs, are better suited to ensure compliance and cooperation than officials elected to representative legislatures. Specifically, Baldwin argues that the long-time horizon of the chiefs tenure in ruling a community encourages cooperation. This argument seems to find empirical support in many developing society concerned with public goods provision (Baldwin and Raffler, 2019). Note, the suggested mechanism by which compliance or cooperation is ensured is not reliant on the form of punishment, collective or individualized, but that the commitment to carry out enforcement is ensured.<sup>1</sup>

In general, individual actors are more likely to comply with authority when they are more certain that others are also likely to do so (Scholz *et al.*, 1995; Scholz and Lubell, 1998), whether that

<sup>1</sup>Previous literature from the laboratory documents that also under perfect monitoring the welfare calculations depend crucially on the time horizon. Since in the short run punishment imposes a cost on individuals, societies might initially be faring worse than in the absence of the opportunity to punish, as observed by Dreber *et al.* (2008). However, in the long run the frequency of punishment decreases and social welfare with punishment may be higher (Gächter *et al.*, 2008).

joint compliance is ensured by a central authority committed to enforcing any rules or individual actors committing themselves to be sanctioned before they even make their choice of whether to cooperate. Authorities and individual members of society often rely on pre-committed sanctions to save on decision costs (Sunstein and Ullmann-Margalit, 1999) but also for peer-to-peer punishment specifically, commitment devices are helping the effects of sanctioning. Ostrom *et al.* (1992) and Balliet (2010) find peer-to-peer punishment most effective when people were able to communicate beforehand indicating that a collective norm was established ex-ante. To a similar effect, when a norm of compliance is perceived as popular or when conformity is enforced, compliance increases (Centola *et al.*, 2005; Willer *et al.*, 2009). Individuals are willing to sustain mutual cooperation conditionally on the expectations that others will cooperate as well (Yamagishi, 1986; Hayashi *et al.*, 1999).

With commitment, individual members of society are now reassured that punishment will be carried out for sure and that they all hold the same standard. At least two consequences emerge from this collectiveness: First, it ensures that the second-order free-riding problem, by which egoists let others shoulder the punishment costs, does not arise (Oliver, 1980; Heckathorn, 1989). Second, as we argue here, when committed punishment is chosen as sanctioning device, the decision individuals face is driven by collective incentives which transform the underlying collective action problem, such as a public goods game, into a coordination game. This means that a game with one equilibrium, nobody contributes, turns into a game with multiple equilibria with one where everybody contributes. Such equilibrium provides a focal point which allows participants to coordinate on high contribution in the induced game.<sup>2</sup>

In this coordination game, selfish individuals might prefer to contribute if this avoids collective punishment being imposed. Otherwise, they prefer to not contribute to the public good. Provided that collective punishment is conditioned on high joint contribution levels qua commitment, the resulting coordination game features, among other equilibria, an equilibrium where everybody contributes and an equilibrium where nobody contributes. Commitment to collective punishment opens the opportunity for a configuration where everyone contributes, solving the underlying collective action problem. Collective incentive schemes alone involve a commitment problem: once a team has produced a surplus, they have no incentive to destroy it or give it away (Holmström, 1982). Thus, in the absence of commitment, the efficacy of collective sanctions hinges on the presence of individuals who are willing to sacrifice own payoffs for punishment. However, if members of a group can commit in advance to a mechanism that punishes everyone in case contributions are below a target, then they can effectively decide in advance whether collective action should take place under incentives that turn it into a public goods game or a coordination game.

In this paper, we test which characteristic of the sanctioning regime ensures cooperation in collective action problems. Following the discussion above, we argue that

*the combination of collective punishment and commitment sustains highest levels of cooperation (Hypothesis).*<sup>3</sup>

To test our hypothesis, we model a central authority in the laboratory by selecting a subject at random to determine punishment, in line with the literature (see e.g., Dickson *et al.*, 2009, 2015) and juxtapose such authority with decentralized sanctioning represented by peer-to-peer punishment or no punishment at all. We then allow for collective punishment either by the central authority or through peers committing to sanctioning when the joint level of cooperation does not surpass a given threshold. In this way, we arrive at a  $2 \times 2$  experimental design, with an

<sup>2</sup>If collective punishment is conditioned on not reaching the full contribution level of the public good, then the resulting coordination game is strategically similar to a minimum effort coordination game (Van Huyck *et al.*, 1990). For a discussion of experimental studies of coordination games, see Camerer (2003), chapter 7.

<sup>3</sup>The experiment was not pre-registered. All treatments conducted within the research agenda are presented in the main text or the online appendix and no observations are omitted.

additional no punishment baseline treatment condition, varying the target of punishment, individuals or groups, and the existence of commitment to sanction.

## 2. Public good games with punishment and imperfect monitoring

Questions of cooperation in collective action problems and pro-social behavior are frequently modeled as public goods game when investigating behavior of politicians or voters more specifically or any members of society in general (see e.g., Ostrom *et al.*, 1992; Baldassarri and Grossman, 2011; Hamman *et al.*, 2011; Butler and Kousser, 2015). We consider an  $n$ -player public goods game where each player holds an endowment of  $e$  units. Each agent  $i$  can either contribute her entire endowment to the public good or not contribute at all,  $g_i \in \{0, e\}$ .<sup>4</sup> We denote the sum of contributions in a group by  $G = \sum_{i=1}^n g_i$ . Each agent  $i$  observes a noisy signal  $s_{ij} \in \{0, e\} = S$  on the true contribution level  $g_j$  of each other agent  $j \neq i$ .<sup>5</sup> This signal reveals the true contribution level with a signal accuracy of  $\Pr(s_{ij} = g_j | g_j = g) \in (0.5, 1]$ . By contrast, the overall level of contributions  $G$  is observed perfectly. We consider the case of a linear return function which implies that the payoff of  $i$  is given by

$$u_i^{PG}(g_i, g_{-i}) = \alpha G + (e - g_i), \quad (1)$$

where  $\alpha$ , with  $\frac{1}{n} < \alpha < 1$ , captures the marginal per capita return from contributing to the public good. If players' utility functions are given by  $u_i^{PG}$  then in the only Nash equilibrium of this game no player contributes, yielding payoffs of  $e$  to everybody.<sup>6</sup> On the other hand, welfare maximization dictates contribution levels of  $e$  for all participants, resulting in payoffs of  $\alpha ne$ .

Testing our hypothesis means demonstrating which punishment regimes lead participants to contribute to the public good and potentially achieve high levels of welfare. To this end, we discuss and analyze several such punishment regimes added to a public goods game. Some of these punishment regimes retain the nature of the public goods game with the equilibrium prediction of no contribution, while some might, depending on the participants' choices, instantiate a coordination game with a no contribution and, among others, a high contribution equilibrium. In this section, we focus on the predictions of the rational choice benchmark for each of these schemes. This is done to give the reader a clear sense of the incentives participants face in the different regimes we consider. Empirically, though, as laid out above, we frequently observe non-zero levels of contributions in public goods games even when the equilibrium prediction calls for no contribution, making a comparison of behavior across punishment regimes an empirically meaningful endeavor. Throughout, we illustrate how deviations from this benchmark may lead to different behavioral predictions and link it to other research which has established further explanations for the choices observed in the discussion.

### 2.1 Collective punishment

Our primary focus is on collective punishment, i.e., forms of punishment where all members of a group are subject to punishment.<sup>7</sup> Whenever a group is collectively punished,  $P$  points are subtracted from all members. In addition and in line with the literature, punishment is costly and its costs are given by  $\beta P$  for some  $\beta \in \mathbb{R}$ , where typically  $\beta < 1$ . We assume that all members of a

<sup>4</sup>Binary contributions allow for a fairly straightforward and for participants comprehensible implementation of imperfect monitoring (Ambrus and Greiner, 2012).

<sup>5</sup>The level of noise was chosen to ensure that there is no symmetric cooperate equilibria in the individual-punishment case.

<sup>6</sup>Naturally, if players have conditionally cooperative preferences and are sufficiently motivated by negative reciprocity and the game is played more than once then cooperation may constitute an equilibrium, at least in early periods.

<sup>7</sup>Alternatively, one can think of punishment mechanisms where subsets of agents are randomly singled out for punishment. Provided the expected punishment is the same the analysis does not change for risk neutral decision makers. However, risk aversion may well amplify the effect of punishing randomly selected agents.

group share the cost of punishment equally, so that in case of collective punishment each group members' payoff is lowered by  $P(1 + \beta)$ . Due to their supposedly different properties, in particular with respect to whether individuals see them as legitimate, we focus on two different scenarios, depending on whether the decision to collectively punish is taken before or after the public goods game has been played. This corresponds to whether there is a possibility to commit to collective punishment.

2.1.1 Collective punishment without commitment

Suppose the decision whether to collectively punish is taken after the public goods game has been played. One agent is drawn at random and decides whether to punish the group.<sup>8</sup> We denote the chosen agent by  $j$  and let  $p_j^C \in \{0, 1\}$  denote her punishment choice. This gives rise to the following payoff function

$$u_i^C(g_i, g_{-i}) = \alpha G + (e - g_i) - p_j^C P(1 + \beta).$$

As the punishment decision is taken after the contribution stage, player  $j$  will decide not to punish and thus in the only sub-game perfect equilibrium no player will contribute.

If players are motivated by negative reciprocity (i.e., players are not entirely selfish) or if the players or the players perceive the interaction to be indefinitely repeated rather than finite (i.e., players do not have a correct understanding of the game) then there may exist equilibria with full cooperation.<sup>9</sup>

2.1.2 Collective punishment with commitment

In this scenario, there is an institution that allows players to commit to collective punishment before the public goods game. When participants have committed to collective punishment, punishment is automatic and contingent on the overall contributions to the public good. More precisely, everybody will be punished in case the sum of contributions falls short of an exogenously set target level  $\bar{G}$ . Otherwise, there is no punishment. This is formalized by the indicator function  $\mathbb{1}[G < \bar{G}]$ . As above, one randomly selected agent may decide whether collective punishment is implemented.<sup>10</sup> We denote her choice by  $p_j^{C-Comm} \in \{0, 1\}$ . This gives rise to the following payoff function

$$u_i^{C-Comm}(g_i, g_{-i}) = \alpha G + (e - g_i) - p_j^{C-Comm} \mathbb{1}[G < \bar{G}] P(1 + \beta).$$

We restrict attention to the case where  $\bar{G} = ne$ , so that a group is punished if one or more members did not contribute.<sup>11</sup>

Note that if player  $j$  decides not to introduce committed collective punishment,  $p_j^{C-Comm} = 0$ , the game reverts to the initial public goods game where all players choose  $g_i = 0$  in equilibrium. If committed collective punishment is introduced,  $p_j^{C-Comm} = 1$ , and provided punishment and its associated cost exceed the net benefit of not contributing,  $(1 + \beta)P \geq (1 - \alpha)e$ , a group members best response is to keep her contributions in line with the group member with the lowest contribution level. Thus, the profile where everybody contributes and the profile where nobody

<sup>8</sup>Results generalize to other collective decision-making mechanisms (i.e., majority vote).

<sup>9</sup>As we show in online appendix A.2, the conditions for cooperation are nevertheless more favorable under collective punishment with commitment than under collective punishment without commitment.

<sup>10</sup>Such procedure models sincere voting, avoiding the influence of strategic considerations often present in vote choice, and implies that individuals are more likely to be pivotal than in large-scale elections but still not influential in all decisions (for the use of this procedure, see Morton and Ou, 2015, 2019).

<sup>11</sup>More lenient punishment institutions may feature lower target levels,  $\bar{G} < ne$ . Such institutions may however only partially overcome the free rider problem as a subset of  $\lfloor \bar{G}/e \rfloor - n$  agents prefers not to contribute.



contributes are both Nash equilibria. The payoffs in these states are given by  $\alpha ne$  and  $e - p(1 + \beta)$ , respectively. The socially optimal payoff is thus achieved in one of this sub-game's equilibria.

Given the equilibria in the sub-games, there are two sub-game perfect equilibria in the induced extensive form game: one where player  $j$  decides against the institution and nobody contributes in the public goods game arm (and nobody contributes in the hypothetical coordination game arm) and one where the institution is formed and everybody contributes in the coordination game arm (and nobody contributes in the hypothetical public goods game arm). There does not exist a sub-game perfect equilibrium where collective punishment is implemented and players do not contribute in the implied coordination game. Thus, the presence of a collective punishment institution may act as a focal point that provides a rationale for the payoff dominant equilibrium of the coordination game.

## 2.2 Peer-to-peer punishment

We benchmark our collective punishment regime against the standard peer-to-peer punishment regimes (Ostrom *et al.*, 1992; Fehr and Gächter, 2000, 2002; Gächter *et al.*, 2008) in addition to a condition without punishment. In noisy environments, these peer-to-peer punishment regimes have a number of downsides, most notably that they are very prone to both type I and type II errors, i.e., punishing contributors or not punishing non-contributors. It is for this reason that collective punishment mechanisms seem particularly appealing in such high-noise environments.

### 2.2.1 Peer-to-peer punishment without commitment

Under this sanctioning regime, after the public good each player  $i$  can decide whether to punish each player  $j$ ,  $p_{ij}^S \in \{0, 1\}$ . In case of punishment,  $P$  points will be subtracted from the punished, resulting in a cost of  $\beta P$  to the punisher. Under peer-to-peer punishment, the payoff function is given by,

$$u_i^S(g_i, g_{-i}) = \alpha G + (e - g_i) - P \sum_{j \neq i}^n p_{ji}^S - \beta P \sum_{j \neq i}^n p_{ij}^S.$$

Since punishment is costly, no rational and selfish player will engage in it. This observation holds regardless of whether information about contributions is noisy. Thus, in the only sub-game perfect equilibrium of this game, no player will be punished and no player will contribute. Of course, empirically, we have seen that people do sometimes punish and that punishment can be effective in sustaining contributions in environments without noise or in low-noise environments (Fehr and Gächter, 2002). In environments characterized by high-noise levels, this punishment regime has been much less effective empirically, as shown in, e.g., Ambrus and Greiner (2012).

As noted in the above in the context of collective punishment without commitment, if players are motivated by negative reciprocity or if the players or the players perceive the interaction to be indefinitely repeated rather than finite then there may exist equilibria with full cooperation.<sup>12</sup>

### 2.2.2 Peer-to-peer punishment with commitment

With this punishment technology, before the public goods game, each player  $i$  announces a contingent plan under which circumstances each other player is punished. When choosing their contribution levels in the public goods game, agents know the punishment plan of all agents. Once signals on contribution levels have been received, punishment is automatic. We use  $p_{ij}^{S-Comm} \in \{0, 1\}$  to denote whether player  $i$  has committed to punish player  $j$ . Punishment is contingent on the noisy signal received by player  $i$  on  $j$ 's contributions,  $s_{ij}$ . When  $i$  commits to

<sup>12</sup>Still, even under these alternative assumptions, the conditions for cooperation are still more favorable under collective punishment with commitment than under peer-to-peer punishment without commitment (see online appendix A.2).

punish  $j$ , the indicator function  $\mathbb{1}[s_{ij} < e]$  captures that  $j$  is punished whenever  $i$  receives the signal that  $j$  did not contribute. The payoff function in the case of peer-to-peer punishment under commitment is, thus, given by

$$u_i^{S-Comm}(g_i, g_{-i}) = \alpha G + (e - g_i) - P \sum_{j \neq i}^n p_{ji}^{S-Comm} \mathbb{1}[s_{ji} < e] - \beta P \sum_{j \neq i}^n p_{ij}^{S-Comm} \mathbb{1}[s_{ij} < e].$$

For high signal accuracy, there exist sub-game perfect Nash equilibria where (at least some) agents commit to punish and all agents contribute in the public good game. Note that for positive noise levels, these equilibria will involve punishment of contributors. Conversely, for low signal accuracy, there do not exist sub-game perfect equilibria where agents commit to punish or contribute in the public good game. The reason behind this is that even for a high committed level of punishment, the payoff from not contributing exceeds the payoff from contributing as contributors and non-contributors are almost observationally equivalent.<sup>13</sup>

### 3. Experimental design

Our experiment featured five main treatments. Our baseline treatment (N) studies imperfect monitoring without the possibility to punish others. The four different punishment mechanisms described in the previous section correspond to a  $2 \times 2$  factorial design. One dimension is defined by whether there is peer-to-peer punishment or collective punishment. The other dimension is defined by whether participants can commit to punishments.<sup>14</sup>

Our main treatments are summarized in Table 1. In all treatments, participants interacted in a 50-period repeated public good game. The reason to choose such a long horizon is to allow for learning and hence to allow us to observe mature decisions of participants once they have become familiar with the environment they make decisions in.

Subjects were randomly matched in groups of four which remained constant for the entire duration of the experiment. In this way, we obtain as many independent observations as there are groups in the experiment. In each period of the public goods game, each participant received an endowment of 20 tokens and could decide to either contribute all of her endowments to a group account or not contribute at all. In order to mitigate the effects of possibly excessive punishment, and to make sure that participants did not leave the experiment with negative earnings, in each period each agent additionally received a payment of 10 tokens which could not be contributed. In addition, the minimum payoff per round was set at zero.<sup>15</sup> While endowments that were kept by agents only benefited themselves, endowments that were contributed to the group account benefited each agent by 10 tokens. Thus, the payoff function in the public good game was given by (1) with  $\alpha = 0.5$  and  $e = 20$ .

All of our treatments featured imperfect monitoring of actions. After the public good stage, each player  $i$  received a noisy signal  $s_{ij}$  on the true contribution level of player  $j$ . Signals were independently distributed across players, implying that two players may receive different signals on the behavior of a third player. With probability 0.6, the signal reflected the true behavior of a participant.<sup>16</sup> With the remaining probability 0.4, a contributor was labeled as a non-contributor and

<sup>13</sup>See online appendix A.1 for a formal proof of this statement. It remains true even if players are motivated by negative reciprocity or misperceive the game as indefinitely repeated, as discussed in online appendix A.2.

<sup>14</sup>In addition, we implemented a treatment with a strong punishment technology, see Ambrus and Greiner (2012), which we discuss in Section E.1 in the online appendix. In addition to the treatments featured here, we ran treatments with lower noise rates and slightly different information structures. See online appendix E.

<sup>15</sup>This limited liability constraint was not reached a single time in our experiment.

<sup>16</sup>For high signal accuracy/low-noise environments, we found that peer-to-peer punishment was successful in solving the free-rider problem. As we are interested in situations where peer-to-peer punishment fails to do so, we focus on low signal accuracy/high-noise levels. See Appendix E.

**Table 1.** Main treatments

	No commitment	Commitment
Peer-to-peer punishment	<b>S</b> (2400,12,48)	<b>S-Comm</b> (4600,23,92)
Collective punishment	<b>C</b> (2600,13,52)	<b>C-Comm</b> (4600,23,92)
No punishment		<b>N</b> (1600,8,32)

Note: The table shows the main treatments. In brackets number of observations, number of 4-player groups, and number of players.

a non-contributor was labeled as a contributor. Hence, signals are informative but there is a substantial chance both that a contributor appears to be a non-contributor and vice versa. Note, for a signal precision of 0.6, there does not exist a symmetric sub-game perfect equilibrium in which everyone contributes.

In addition, in all treatments, participants were correctly informed about the sum of contributions to the group account,  $G$ . Subjects were made aware of this information structure in the instructions and on the feedback screens. Our choice of imperfect monitoring of contributions but perfect monitoring of the sum of contribution levels is motivated by the observation that individual efforts are often observed in a more noisy way than the results of a collaborative effort. While the information participants receive is identical in all treatments, and therefore none of the technologies can be accused of being more likely to present false positives when it comes to identifying non-contributors than another, treatments exclusively differ in the punishment technology available to participants.

In the no punishment treatment, **N** participants simply played the contribution game without a punishment stage. In the punishment treatments without commitment, participants could decide on punishing other participants after the contribution stage where they received a noisy signal of individual choices but were perfectly informed about total contributions. In the standard peer-to-peer punishment treatment **S**, each participant  $i$  was asked whether she wanted to subtract  $p = 15$  punishment points from each other participant  $j$  or not. Punishment costs are set at  $\beta = \frac{1}{3}$ , implying that punishing another player costs 5 tokens. After the punishment stage, the punishment points received and the cost for punishment of others were subtracted from the earnings in the contribution stage. Afterwards, the final payoffs for a round were presented to participants.

In the collective punishment treatment **C**, each participant was asked, again after receiving a noisy signal on the contributions and being informed about total contributions, whether she would like to subtract 15 points from everybody (including herself) at a cost of 5 to everybody.<sup>17</sup> Subjects were made aware that the decision of one of the players would be chosen at random for implementation.<sup>18</sup> The fact that everybody had to bear the cost of punishment allows us to directly compare collective to peer-to-peer punishment.

In the punishment treatments with commitment, participants could commit to punishing other participants before the contribution stage. Subjects were informed about the punishment decisions of others at the contribution stage and punishment was carried out automatically given the information received from the contribution stage. In the peer-to-peer punishment with commitment treatment **S-Comm**, each participant was asked whether they commit to

<sup>17</sup>Thus, under collective punishment, 20 points are subtracted from everybody. We have chosen to represent this in terms of a cost (5 points) and a punishment (15 points) to ensure comparability of the results to our other treatments and the previous literature.

<sup>18</sup>We have chosen this collective decision-making rule, as it (i) yields a higher number of observations of punishment decisions than ex-ante chosen decision makers in each round, (ii) empowers all participants as compared to one constant decision maker, (iii) avoids the tie splitting problem with four participants, and (iv) avoids certain indifference cases as compared to majority vote.

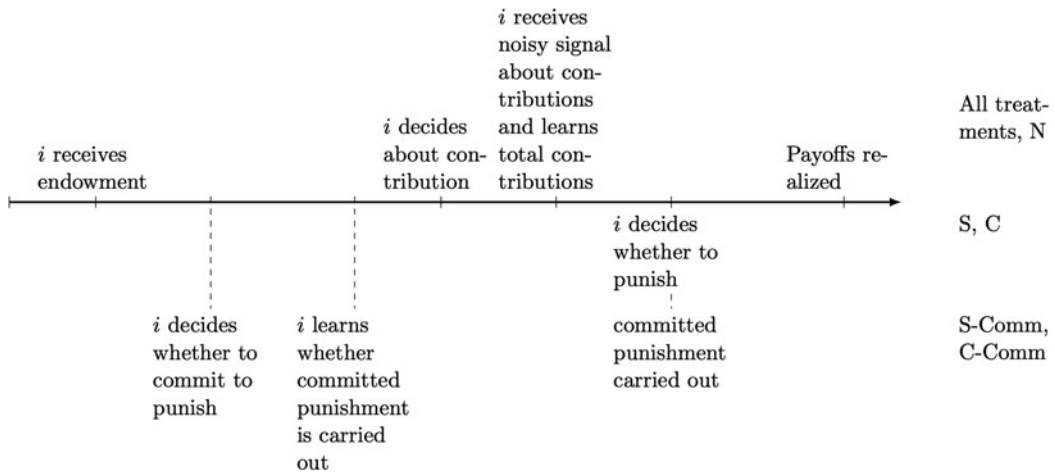


Figure 1. Sequence of moves within one round of the public goods game by treatment.

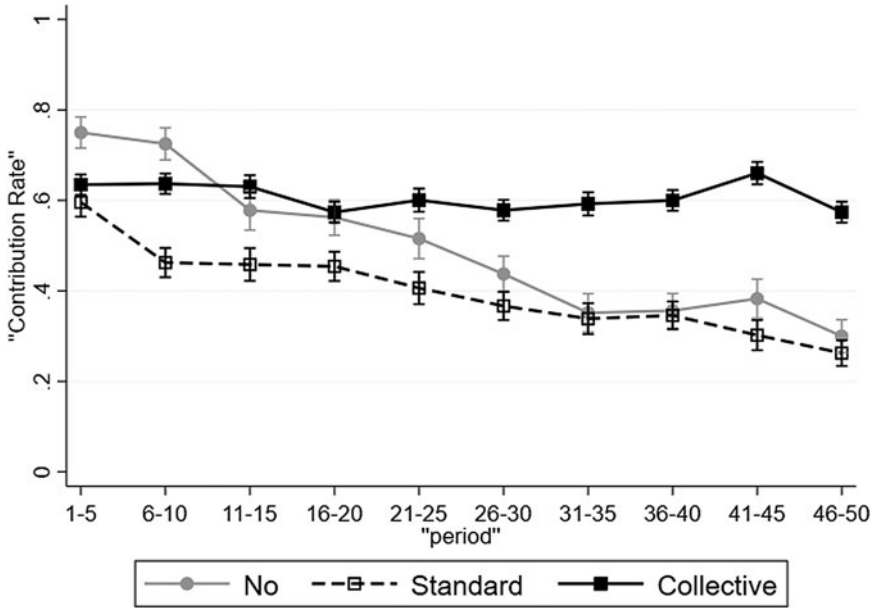
punish another participant if they receive the noisy signal that (i) the participant contributed or (ii) did not contribute. Thus, each participant had to make six decisions at this stage.

In the collective punishment with commitment treatment **C-Comm**, participants were asked whether they commit to punishing everybody (including themselves) in case somebody did not contribute. Before the contribution stage, the choice of one participant was randomly implemented and everybody was made aware whether a collective punishment mechanism was in place.<sup>19</sup> Figure 1 below summarizes the sequence of moves shared by all treatments and those steps that are specific to only some or just one of the treatments.

Quantifying the effect of punishment, we compare the standard peer-to-peer punishment (S) to the no punishment treatment (N). The effect of collective punishment as well as of the ability to commit to punishment is then estimated by (1) comparing standard peer-to-peer punishment (S) to collective punishment (C); and standard peer-to-peer punishment (S) to peer-to-peer punishment featuring the ability to commit to punishment (S-Comm) as well as to collective punishment with commitment (C-Comm). Measuring whether individuals choose to turn the public goods game into a coordination game by committing to punishment, we further contrast cooperation and payoffs when individual participants choose no commitment to when they do (in treatment C-Comm). Finally, and most importantly, we identify the joint effect of commitment and collective punishment by juxtaposing cooperation and payoffs across treatments S, S-Comm, C, and C-Comm.

We report the estimated regression models throughout the results section. Since we do not expect equilibrium predictions to matter until participants have had time to learn and adapt their behavior, we run our analysis on the last 10 periods, in addition to on the full set of rounds, as it is common practice in experimental tests of public goods or coordination games; we argue that behavior in those last periods is most stable and we run hypothesis tests over treatment effects on this subset of data. For assessing learning, we further compare the first half or the second half of periods separately, show results for the last 10 rounds only, and show regression models including a linear time trend. We analyze the data using linear panel regression

<sup>19</sup>Collective punishment only conditions on whether the sum of contributions is maximal and thus requires even less information than is available here. Our formulation is just one possible form of collective punishment and one may think of alternative forms that are more forgiving in the sense that only a fraction of individuals is required to contribute. Such rules may however be subject to coordination problems as it is not clear who will free ride.



**Figure 2.** Contribution rates in the cases of no punishment (“No”), standard peer-to-peer punishment (“Standard,” S), and collective punishment with commitment (“Collective,” C-COMM).

models which allow us to cluster standard errors at the group level and to account for auto-correlation.<sup>20</sup>

## 4. Results

### 4.1 Efficacy of collective punishment with commitment

In a first step, we assess the performance of collective punishment with commitment. We focus on the collective punishment treatment *with* commitment as a benchmark for testing our hypothesis. We first compare contribution and payoff rates under collective punishment with commitment to the two natural benchmark scenarios of standard peer-to-peer punishment (treatment S) and the case where the possibility of punishment is absent (treatment N). Those are the benchmark cases most prominently studied in the literature and most frequently observed in the field. In Section 4.3, we then include our other treatment variations in order to gauge the relative importance of the effects of commitment and collectiveness.

Figure 2 plots contribution rates over time for the three treatments under consideration. While cooperation levels appear to be fairly high and stable over time under collective punishment with commitment, they are decreasing over time in the other two scenarios; more than halving over the course of the experiment.

To compare the effects on cooperation/contribution, we run the following regression.

$$C_i^t = \alpha + \beta_1 \delta_S + \beta_2 \delta_{C-Comm} + \epsilon_i^t \tag{2}$$

Here the dependent variable  $C_i^t$  is a dummy for whether player  $i$  contributes in period  $t$ , and the independent variables  $\delta_S$  and  $\delta_{C-Comm}$  are dummy variables for treatments S and C-Comm, respectively. The constant  $\alpha$  captures the baseline effect in the case of no punishment.

<sup>20</sup>Whenever we report a finding as significant, we either report the level of significance applied or the exact p-value. Giving a fuller account of uncertainty in our results this way should enable readers to evaluate better the robustness of our results.

Standard errors are clustered at the matching group level. Table 2 shows the results. The columns differ according to how many periods are taken into account in the regression. Column (2) differs from the other regressions in that it includes a linear time trend. Column (5) presents the results for the last 10 periods upon which much of the analysis draws.

It turns out that collective punishment with commitment has a positive effect on cooperation, relative to the baseline of no punishment, and this effect is significant at the 1 percent-level. In contrast, peer-to-peer punishment leads to lower cooperation than the baseline of no punishment, though this effect is not statistically significant. A test of whether the coefficients for S and C-Comm are equal confirms that collective punishment with commitment yields higher cooperation than peer-to-peer punishment. Column (2) in Table 2 shows that contributions are decreasing over time in the no punishment scenario. Further post-regression tests reveal that this is also the case under peer-to-peer punishment ( $\beta_3 + \beta_4 \approx -0.06, p = 0.00$ ) while the time trend for collective punishment with commitment is not significantly different from zero ( $\beta_3 + \beta_5 \approx 0.00, p = 0.70$ ). Summarizing our findings so far we have:

Result 1: Contribution levels are higher under collective punishment with commitment compared to the no punishment scenario or to peer-to-peer punishment. There is no difference between the no punishment scenario and peer-to-peer punishment. While contribution levels are stable under collective punishment with commitment, they are decreasing under peer-to-peer punishment and in the no punishment scenario.

Collective punishment supports relatively high contribution levels over time. We now move on to assess its relevance for overall welfare; taking into account the welfare loss incurred if punishments are executed. The evolution of net profits over time for the different treatments is displayed in Figure 3.

Table 3 reports the results of the following regression

$$\pi_i^t = \alpha + \beta_1 \delta_S + \beta_2 \delta_{C-Comm} + \epsilon_i^t \tag{3}$$

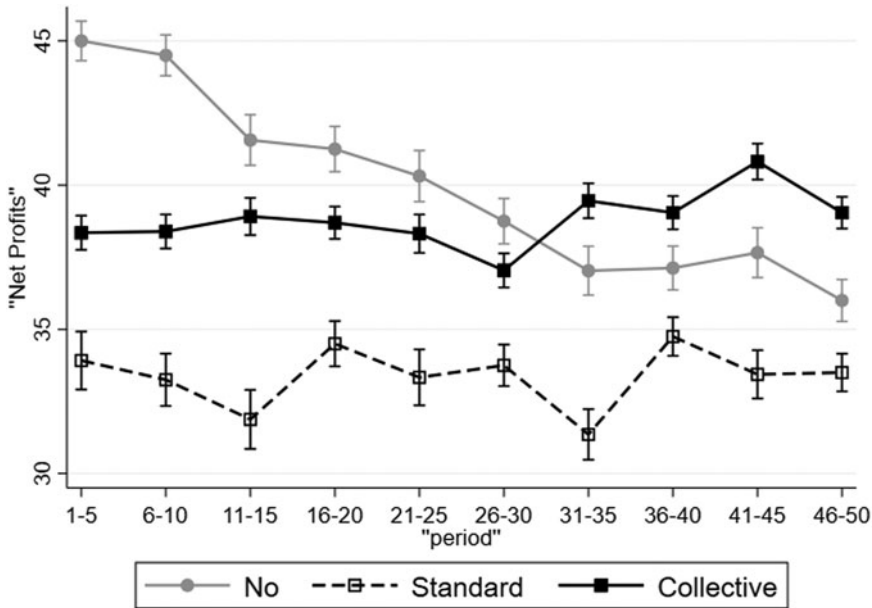
where  $\pi_i$  are player  $i$ 's net profits at  $t$ , i.e., their earnings from the public good game net off any costs incurred for punishment or being punished.

We find that collective punishment with commitment has a positive effect on payoffs relative to the baseline of no punishment, this effect is significant at the 10 percent-level. Peer-to-peer

Table 2. OLS estimates of regression equation (2)

Variables	(1) All periods	(2) Time trend	(3) 1st Half	(4) 2nd Half	(5) Last 10 periods
S ( $\beta_1$ )	-0.095 (0.087)	-0.202** (0.084)	-0.154* (0.086)	-0.036 (0.094)	-0.053 (0.105)
C-Comm ( $\beta_2$ )	0.107 (0.064)	-0.141* (0.074)	-0.023 (0.068)	0.238*** (0.068)	0.279*** (0.077)
period ( $\beta_3$ )		-0.010*** (0.001)			
period × S ( $\beta_4$ )		0.004** (0.001)			
period × C-Comm ( $\beta_5$ )		0.009*** (0.001)			
Constant ( $\alpha$ )	0.500*** (0.045)	0.761*** (0.060)	0.635*** (0.054)	0.365*** (0.041)	0.341*** (0.052)
p-value Test $\beta_1 = \beta_2$	0.025**		0.1074	0.009***	0.003***
Observations	8600	8600	4300	4300	1720
R <sup>2</sup>	0.031	0.056	0.016	0.068	0.098

Note: LPM estimates of cooperation regressed on treatment dummies (equation (2)). Robust standard errors clustered at the matching group level are in parenthesis. \*\*\*p < 0.01, \*\*p < 0.05, \*p < 0.1.



**Figure 3.** Profits in the cases of no punishment (“No”), peer-to-peer punishment (“Standard”, S), and collective punishment with commitment (“Collective”, C-COMM).

punishment is associated with lower payoffs than the baseline of no punishment, although this effect is not statistically significant across the last 10 periods. However, collective punishment with commitment yields significantly higher profit than peer-to-peer punishment without commitment, as revealed by the *t*-test of equality of coefficients  $\beta_1$  and  $\beta_2$ . The effect is substantial and corresponds to an increase of 6.585 ECU, or roughly 20 percent. Studying time trends reveals that profits in the no punishment case are decreasing over time. Additional post-regression tests confirm that there are no significant time trends in payoff rates under peer-to-peer punishment ( $\beta_3 +$

**Table 3.** OLS estimates of regression equation ((3))

Variables	(1) All periods	(2) Time trend	(3) 1st Half	(4) 2nd Half	(5) Last 10 periods
S ( $\beta_1$ )	-6.508*** (1.979)	-11.790*** (2.352)	-9.117*** (2.127)	-3.900* (2.099)	-3.354 (2.350)
C-Comm ( $\beta_2$ )	-1.209 (1.567)	-7.271*** (1.843)	-4.257*** (1.650)	1.839 (1.665)	3.231* (1.856)
period ( $\beta_3$ )		-0.205*** (0.026)			
period × S ( $\beta_4$ )		0.207*** (0.057)			
period × C-Comm ( $\beta_5$ )		0.238*** (0.041)			
Constant ( $\alpha$ )	40*** (0.906)	45.22*** (1.200)	42.70*** (1.082)	37.30*** (0.828)	36.81*** (1.043)
p-value Test $\beta_1 = \beta_2$	0.0192**	0.0733*	0.0337**	0.0218**	0.0154**
Observations	8600	8600	4300	4300	1720
R <sup>2</sup>	0.043	0.055	0.059	0.044	0.061

Note: OLS regression with profits regressed on treatment dummies (equation (3)). Robust standard errors clustered at the matching group level are in parenthesis. \*\*\*p < 0.01, \*\*p < 0.05, \*p < 0.1.

$\beta_4 \approx 0.00$ ,  $p = 0.97$ ). While payoffs seem to be increasing under collective punishment, this effect is not significant ( $\beta_3 + \beta_5 \approx 0.03$ ,  $p = 0.31$ ), i.e., we cannot reject  $\beta_3 + \beta_5 = 0$ .

**Result 2:** Payoff levels under collective punishment with commitment are higher than in the no punishment scenario and are higher than under peer-to-peer punishment. Payoff levels in the no punishment case are not significantly higher than under peer-to-peer punishment. While payoff levels are decreasing in the no punishment scenario, they are stable under collective punishment with commitment and under peer-to-peer punishment without commitment.

Collective punishment restores relatively high levels of welfare in a noisy environment where peer-to-peer punishment does not work as well as in the noiseless case. While positive payoff implications of collective punishment with commitment (as compared to the no-punishment case) start to become evident only toward the end of the experiment, the pattern of time trends documented in column (2) of Tables 2 and 3 points toward long-term positive welfare effects. This interpretation is reinforced by focusing on groups where collective punishment is committed.

#### 4.2 Collective punishment as a coordination device

As previously noted, whenever collective punishment is committed participants face a coordination game with two Nash equilibria: in one nobody contributes and in the other everybody contributes. By contrast, when no collective punishment is committed the resulting sub-game corresponds to the original public goods game. By choosing whether to commit to implementing collective punishment agents may decide whether to play the normal public goods game or a coordination game. In the latter case, the resulting extensive form game has two sub-game perfect Nash equilibria: in one collective punishment is implemented and everybody contributes whereas in the other collective punishment is not implemented and nobody contributes (see Section 2). There is no sub-game perfect equilibrium where collective punishment is implemented and the no contribution equilibrium of the coordination game is played. The presence of collective punishment acts as a focal point in the implied coordination game.

In order to assess this prediction, in the C-Comm treatment, we now contrast contribution rates in groups where collective punishment is committed to those where it is not committed. Figure 4 plots contribution rates in these two cases over the course of the experiment and includes the no punishment (N) treatment as benchmark.

While behavior in groups in the collective punishment treatment where collective punishment was not implemented closely resembles behavior in the no-punishment treatment, contributions in groups with implemented collective punishment are significantly higher and increasing over the course of the experiment, amounting to 94 percent in the last round.<sup>21</sup>

Analysis of the effect of actually committed collective punishment on cooperation suffers from endogeneity problems, as which groups choose to commit to collective punishment is not exogenous. Typically, we would think of the groups who commit and those who do not as different, in the sense that they have a different distribution of types. We can address this selection problem by exploiting the design feature that, while all group members make a choice on whether they would like to implement collective punishment, only the decision of one randomly selected group member is implemented (see Section 3). We can hence compare groups where the same number of participants opt for collective punishment, but where the random (and exogenous) selection procedure implemented a different decision circumventing issues of post-treatment bias.

<sup>21</sup>This is in stark contrast to previous work on minimum effort games (see, e.g., Van Huyck *et al.*, 1990) where participants were unable to reach high effort equilibria. While one may argue that these experiments feature more complex games with more than two strategies, there is evidence showing that even in simpler two-person  $2 \times 2$ -games coordination on high effort equilibria is fairly demanding (see Battalio *et al.*, 2001).



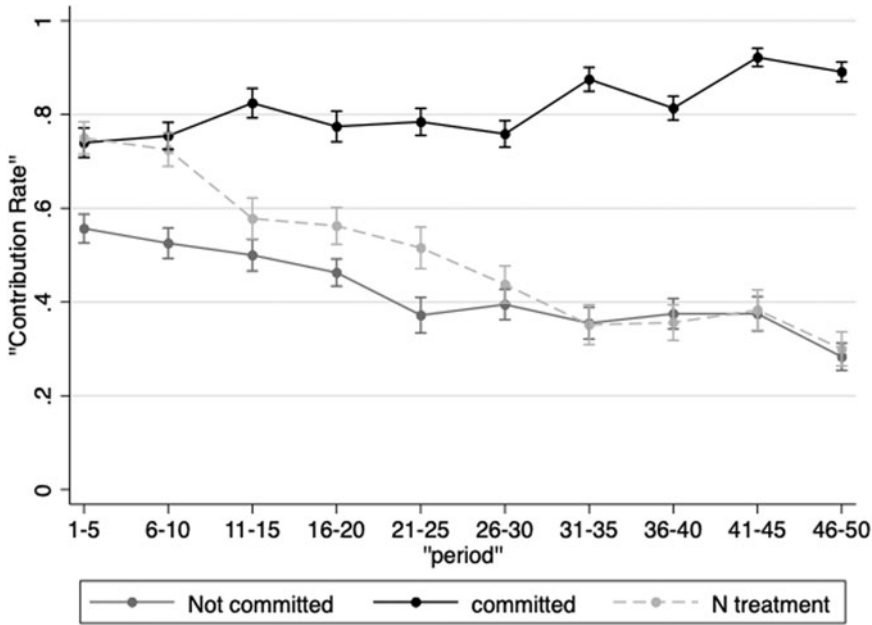


Figure 4. Contributions with and without collective punishment committed.

Table 4 (upper half) shows the results.

Horizontal comparisons illustrate the selection problem. Irrespective of whether collective punishment is implemented or not, there is always more cooperation when more people voted in favor of the institution. Indeed, the correlation between past individual contributions and voting for punishment is statistically significant ( $\rho = 0.1872^{***}$ ) and so is the correlation between group past contributions and the share of group members voting for the punishment technology. These correlations may reflect a number of things. There might be different types in society and those who tend to vote for punishment institutions are also those who are more likely to contribute; or that the institution is successful in sustaining cooperation and people vote for these institutions, the institution successfully keeps contributions high and then people vote for them again.

Vertical comparisons illustrate the effect of the punishment institution being implemented. Conditional on the number of people who voted in favor of the institution (1, 2, or 3), whether it is implemented is exogenous. The table shows that there is always more cooperation if the institution is implemented, though the difference is not statistically significant with one vote in favor ( $t$ -test,  $p = 0.32$ ). It is statistically significant for 2 votes ( $p < 0.01$ ) and 3 votes ( $p = 0.02$ ).

Table 4. Contributions and profits when collective punishment is and is not committed

Votes in favor	0	1	2	3	4	Average
	<i>Contribution rates</i>					
Coll. Pun. committed	–	0.55	0.76	0.92	0.98	0.81
Coll. Pun. not committed	0.33	0.44	0.43	0.56	–	0.43
	<i>Profits</i>					
Coll. Pun. committed	–	33.8	35.1	44.4	48.1	40.5
Coll. Pun. not committed	36.6	37.8	34.6	40.6	–	37.3
Observations	172	331	256	308	83	

Note: The table shows average contribution rates (across all periods) depending on how many group members voted in favor of implementing the collective punishment mechanism and whether or not it was actually implemented.

Result 3: On average, contribution levels are significantly higher when collective punishment is committed compared to the no punishment scenario and compared to the case where collective punishment is not committed.

We now proceed to assess the implications of committed collective punishments for welfare. Figure 5 plots the evolution of payoffs in the two subgroups and contrasts it to the no punishment treatment. While profits of participants in groups where collective punishment is not implemented are statistically indistinguishable from profits made by participants in the no punishment treatment, the profit of participants in groups with implemented collective punishment is clearly higher in the later periods of the experiment. Further, payoffs are evidently increasing for participants in groups where punishment is committed ( $\beta = 0.1635^{***}$ ) and are decreasing for the other two reference groups (non-committed:  $\beta = -0.099^{***}$ ; N:  $\beta = -0.204^{***}$ ). Thus, while payoffs in groups with implemented collective punishment are initially lower than in the no punishment case, this picture is reversed by the end of the experiment. This suggests that it takes time for agents to learn to coordinate on the high contribution equilibrium when punishment is introduced.

Table 4 (lower half) shows payoff comparisons when we condition on the number of group members who voted in favor of establishing the institution. Interestingly, in relatively uncooperative groups where only one person voted in favor of the mechanism, payoffs are lower when the mechanism is implemented (though not statistically different,  $p = 0.42$ ). As the cooperation rate is low in these groups, punishment is relatively often executed, which lowers payoffs. As expected payoffs are higher when the institution is implemented both when two voters are in favor ( $p = 0.06$ ) and when three voters are in favor ( $p = 0.12$ ).

Result 4: On average, payoff levels are higher when collective punishment is committed compared to the no punishment scenario and compared to the case where collective punishment is not committed.

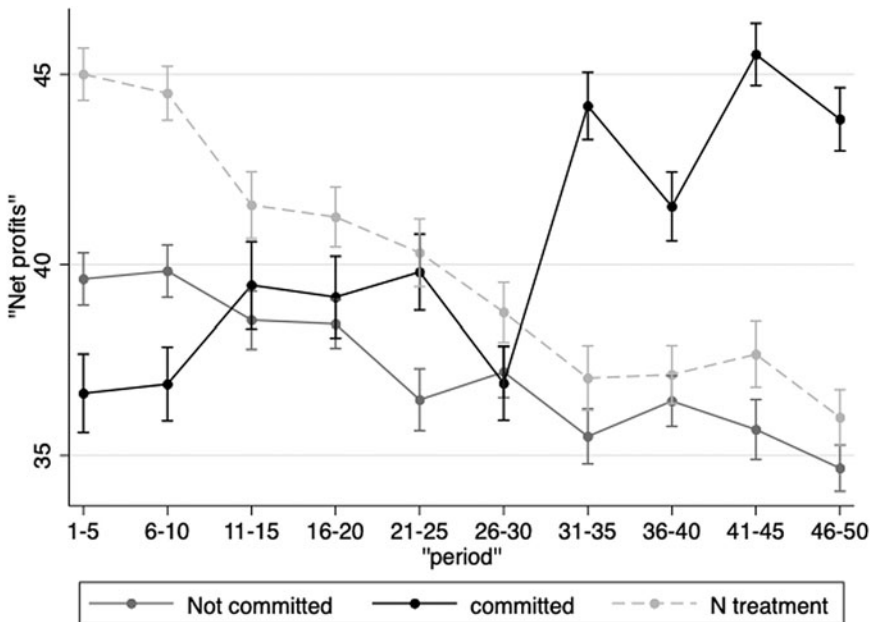


Figure 5. Profits with and without collective punishment committed

**4.3 The role of commitment and collectiveness**

In order to assess the relative importance of commitment and collectiveness for cooperation and profit, we now combine the data from treatments S, C, S-Comm, and C-Comm. To investigate the effect on cooperation, we run the following regression

$$C_i = \alpha + \beta_1 \delta_{Comm} + \beta_2 \delta_{Coll} + \beta_3 (\delta_{Comm} \times \delta_{Coll}) + \epsilon_i \tag{4}$$

Here  $\delta_{Comm}$  is a dummy indicating that the treatment was one with commitment (S-Comm or C-Comm) and  $\delta_{Coll}$  is a dummy indicating that the treatment was one with collective punishment (C or C-Comm).

Table 5 displays the results. As before, the columns differ according to how many periods are taken into account in the regression, and again we focus on the results for the last 10 periods (column (5)). In the baseline case (peer-to-peer punishment without commitment), the average cooperation rate is around 28.7 percent in the last 10 periods. The commitment coefficient is significant and positive, indicating that adding commitment to either peer-to-peer punishment ( $\beta_1$ ) or to collective punishment ( $\beta_1 + \beta_3$ ) has a significant positive effect on the rate of cooperation, increasing it to 49.7 percent in the case of peer-to-peer punishment and even 66.3 percent in the case of collective punishment. By contrast, the collectiveness coefficient ( $\beta_2$ ), though positive, is small and not significant, implying that *without* commitment, collective punishment does not lead to higher contributions. The average cooperation rate in C is 29.7 percent. The sum of the collective coefficient and the interaction term coefficient ( $\beta_2 + \beta_3$ ) shows higher contributions for committed punishment when it is collective.

Result 5: The ability to commit to punishment increases contributions compared to peer-to-peer punishment. While in the absence of commitment, collective punishment does not lead to higher contributions, it leads to weakly higher contributions in the presence of commitment.

Table 5. OLS estimates of regression equation (4)

Variables	(1) All periods	(2) Time trend	(3) 1st Half	(4) 2nd Half	(5) Last 10 periods
commit ( $\beta_1$ )	0.179** (0.085)	0.106 (0.078)	0.138* (0.078)	0.220** (0.099)	0.210* (0.110)
coll ( $\beta_2$ )	0.0204 (0.088)	0.061 (0.080)	0.0307 (0.083)	0.0101 (0.101)	0.0106 (0.109)
coll×commit ( $\beta_3$ )	0.0254 (0.116)	-0.120 (0.119)	-0.0374 (0.111)	0.0882 (0.133)	0.156 (0.146)
period		-0.006*** (0.001)			
period × commit		0.003 (0.002)			
period × coll		-0.002 (0.002)			
period × comm × coll		0.006* (0.003)			
Constant ( $\alpha$ )	0.405*** (0.075)	0.559*** (0.060)	0.481*** (0.067)	0.328*** (0.085)	0.287*** (0.090)
Observations	9800	9800	4900	4900	1960
p-value $\beta_2 + \beta_3$	0.5427	0.5084	0.9291	0.2592	0.0920*
p-value $\beta_1 + \beta_3$	0.0120**	0.8722	0.2073	0.0010***	0.0004***
R <sup>2</sup>	0.038	0.061	0.015	0.075	0.098

Note: LPM estimates of cooperation regressed on treatment dummies (equation (4)). Robust standard errors clustered at the matching group level are in parenthesis. \*\*\*p < 0.01, \*\*p < 0.05, \*p < 0.1.

One might wonder how successful the standard technology is depending on how many people have committed to punishing how many others. To this end, we analyzed contribution rates in treatment S-COMM separately for the cases where fewer than 4 individual punishments are committed in a group, where between 4 and 6 individual punishments are committed, 7–9 punishments, or 10 or more punishments. (The maximum amount of individual punishments committed is  $4 \times 3 = 12$ , the case where all group members decide to punish all others. This case did not occur in the experiment, though.) The cooperation rate is 61 percent (52 percent in the last 10 periods) if fewer than 4 punishments are committed. It is 56 percent (42 percent in the last 10 periods) if between 4 and 6 punishments are committed. It is 52 percent if 7–9 punishments are committed, but there are only 23 observations for this case and none in the last 10 periods. There are no observations for the case of 10 or more committed individual punishments. These cooperation rates are always lower than those under committed collective punishment. This analysis shows that one of the reasons that standard punishment does not work as well as collective punishment, even with commitment, is that under the standard technology individuals do not manage to coordinate on situations where a high level of punishment is overall committed. The more targeted punishments, by contrast, do not seem to raise contribution rates sufficiently.

We now move on to discuss the relative importance of collectiveness and commitment on payoffs (net of punishments and punishment costs). To this end, we ran the following regression

$$\pi_i = \alpha + \beta_1 \delta_{Comm} + \beta_2 \delta_{Coll} + \beta_3 (\delta_{Comm} \times \delta_{Coll}) + \epsilon_i, \quad (5)$$

the results of which are presented in Table 6. Neither the commitment coefficient nor the collectiveness coefficient is now significant. However, both the sum of the commitment and interaction coefficients and the sum of the collectiveness and interaction coefficients are significant. This means that although neither commitment nor collectiveness is sufficient to increase payoffs, they are effective when combined.

**Result 6:** Neither the ability to commit to punishment nor whether punishment is collective or not have an effect on profits (taken on their own). The combination of collective and committed punishment leads to higher payoffs than peer-to-peer punishment with commitment and higher payoffs than collective punishment without commitment.

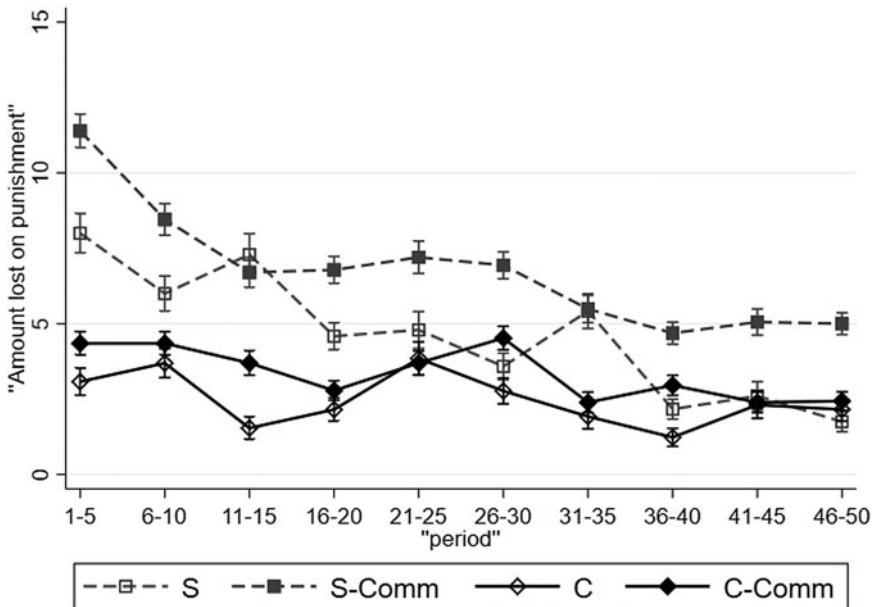
It is interesting to note that while the ability to commit to punishment has a positive effect on contribution levels (as compared to peer-to-peer punishment), only the combination of collective and commitment results in higher payoffs. The main reason for this is that peer-to-peer punishment with commitment results in excessive welfare losses due to punishment, as all too often contributors are incorrectly labeled as non-contributors and consequently punished. This can be inferred from Figure 6 plotting the surplus loss due to punishment (allocated and received). While participants lose approximately 5 points in the last rounds under peer-to-peer punishment with commitment, the surplus loss is approximately half of that in the other treatments. An alternative way of analyzing the surplus loss associated with a punishment technology is to calculate the ratio of net profits (after punishment) to the gross profits (before punishment), thus providing a scaled efficiency measure for a punishment technology. The higher this profit ratio the less wasteful the punishment. This fraction is 0.83 for peer-to-peer punishment (0.86 across the last 10 periods) with commitment, 0.87 for peer-to-peer punishment (0.93, last 10 periods), 0.92 for collective punishment without commitment (0.93, last 10 periods), and 0.93 for collective punishment with commitment (0.94, last 10 periods). A two-sided rank-sum test confirms significant differences (at the 1 percent level) for all pairwise comparisons based on all periods, except the one between the two collective treatments. For the last 10 periods, only the difference between S-COMM and the remaining treatments is statistically significant. Collective punishment

**Table 6.** OLS estimates of regression equation (5)

Variables	(1) All periods	(2) Time trend	(3) 1st Half	(4) 2nd Half	(5) Last 10 periods
commit	0.746 (2.734)	-1.290 (3.396)	-0.358 (2.945)	1.850 (2.933)	1.646 (3.035)
coll	2.485 (1.987)	5.641** (2.390)	3.694* (2.131)	1.277 (2.136)	0.196 (2.359)
coll×commit	2.436 (3.399)	-0.613 (4.25)	1.465 (3.646)	3.406 (3.635)	6.159 (3.818)
period		0.002 (0.050)			
period × commit		0.080 (0.083)			
period × coll		-0.124** (0.061)			
period × commit × coll		0.120 (0.101)			
Constant	33.49*** (1.757)	33.44*** (2.020)	33.58*** (1.828)	33.40*** (1.926)	33.46*** (2.103)
Observations	9800	9800	4900	4900	1960
p-value coll+coll×commit	0.0806*	0.1594	0.0876*	0.1179	0.0395**
p-value comm+coll×commit	0.1216	0.4608	0.6090	0.0181**	0.0015***
R <sup>2</sup>	0.028	0.037	0.027	0.040	0.073

Note: OLS regression with profits regressed on treatment dummies (equation (5)). Robust standard errors clustered at the matching group level are in parenthesis. \*\*\*p < 0.01, \*\*p < 0.05, \*p < 0.1.

(with and without commitment) thus results in less surplus lost due to punishment. This shows that the collective punishment technology is successful in that the threat of punishment contribution is high without causing excessive surplus loss induced by punishment actually being carried out.



**Figure 6.** Surplus loss due to punishment.

## 5. Discussion and conclusion

We have demonstrated that collective sanctions may enable groups to achieve high contribution and welfare levels in an environment where imperfect monitoring makes this inherently difficult. Committed collective punishment induces a coordination game and at the same time provides a rationale for playing the welfare maximizing equilibrium of this game. We find that a key prerequisite for the effectiveness of collective sanctions is the ability to commit to punishment before the collaborative effort. We argue that it is precisely the combination of collectiveness of punishment, influence on the decision to apply collective sanctions, and the insurance that sanctions will be applied that ensures successful collective action, even over a longer period of time where often the amount of free-riding rises.

In very general terms, we argue that a regime with the authority to punish only does so effectively trying to ensure cooperation, when individuals voluntarily select to be governed by that authority. Democracies, featuring the opportunity for (admittedly costly) exit, function better in fulfilling its objectives, e.g., providing public goods, when they enjoy system support. Such support is higher when the institutions are trusted (Wagner *et al.*, 2009), when they are responsive (Kim 2009), or possess fair procedures (Magalhães, 2016). It also has been established that participating in the decision-making process within these institutions, i.e., voting, increases system support (Finkel, 1987; Bowler and Donovan, 2002). We randomly vary the opportunity to participate in the decision how the institution will look like and establish that (1) individuals are willing to submit themselves to such an institution over and over again, a measure of system support, even if it imposes collective sanctions, and (2) the existence of that choice improves group cooperation particularly when collective sanctions are present.

We further speak to the question which institutions enable the government to perform its functions well. In particular, we contribute to characterizing those institutions fostering the provision of public goods and successful collective action. It is said that tackling the complexity of modern societies, representative democracy with an attached ever-growing bureaucracy is ill-equipped to solve the challenges of the day and participation in the process of governing at lower levels—the neighborhood, the work place, the industry, etc.—is seen as providing better outcomes (Fung and Wright, 2001). The World Bank, for example, has been pushing participatory budgeting as standard for implementing its Billion-\$ development programs world wide (Goldfrank, 2012), an institution that seems particularly effective because it tends to stick around (Touchton and Wampler, 2014). Participatory institutions yield better outcomes by creating aware and involved citizens (Agrawal, 2005) but many examples are not benefiting all members of the targeted group often being elite dominated (Mansuri and Rao, 2004), face a trade-off between equity and efficiency (Hong and Cho, 2018), or lack any evidence of improving outcomes at all (Mansuri and Rao, 2012). The institutions we create assign the ability to be involved in creating it, directly testing whether it is the participation in the decision how one is governed, i.e., is collective sanctioning implemented, and whether it is giving everyone the same power over choosing the institution, which yields better government performance. Legal studies, a while ago, have already discovered that “collective sanction incentivizes group members to monitor and marginalize the conduct of conflict entrepreneurs (Drumbl, 2004, p.551).” What our theoretical analysis shows, and the experimental evidence underscores, such incentives only arise when collective action is paired with self-selecting into such regime to avoid a commitment problem.

When we differentiate between peer-to-peer punishment with commitment and collective punishment with commitment, we find that a contract between all members of the group is necessary and not just a contract to allow for high contributions; and, for groups where collective punishment is committed, contributions are even increasing over time, pointing toward beneficial long-term consequences of such institution. Our conclusion that participation of individuals in the decision to apply collective sanctions is also generating high levels of contribution is partially in contrast to findings by Baldwin (2013, 2019) where members of society seem to appreciate the

long-term horizon of non-elected enforcers more than their influence in picking those to carry out sanctioning. However, our results are very much in line with the need for legitimate enforcement that often arises from the way how enforcers are chosen (Dickson *et al.*, 2015, 2009)—where such enforcers seem to be accepted even if the collective sanctioning of all for a rule violation of just one individual seems dictatorial. Such seemingly harsh punishment of everyone is not uncommon outside of the lab; sports fans have known the exclusion from games as punishment for lighting fireworks or invading the pitch, an infringement mostly carried out only by a few. An extension to our finding would also be to consider institutions where rewarding is collective, as frequently observed in sports or at work (Heckathorn, 1988).

Clearly, the specific details of the sanctioning regime we implement determine what aspect of the institution driving behavior we are able to identify; similarly, it sets boundaries on whether and which motivations behind individual's choices we are able to parse. Peer-to-peer punishment and collective punishment, as implemented in the experiment, only differ in the target of imposed sanctions while punishment mechanism with and without commitment differs not only in whether participants bind themselves to punishment but also when they learn about what that punishment would be (before or after they make their decision to contribute). In other words, we cannot separate the effect of a binding decision and of information but consider both to be crucial features of the commitment institution; separating both is left for future research. Keeping the information structure constant across treatments allows us to cleanly identify the effect of the punishment institution on behavior. Variation in choices across groups, then, gives an indication of what may drive the positive effects of committed collective sanctioning: (1) the decision whether to commit to punishment empirically correlates with cooperation, presumably because cooperatively minded people are more likely to build an institution in the first place. (2) Observing how many others are willing to punish, as in the peer-to-peer sanctioning regime with commitment where we find that pre-committed punishment of many or all of the other group members is rare, suggests that individuals are not able to coordinate to widely applied punishment without an institution that commits them to doing so. This is interesting, given that the commitment to imposing sanctions collectively means potentially punishing contributors, which is usually seen as unfair and, in turn, should decrease the appeal of such an institution. It may be that participating in the decision to set-up collective punishment, as a form of realized procedural fairness, seems to be enough to make the threat of collective punishment acceptable; similarly, setting up collective sanctions may be seen as a less involved decision than individuals targeting each other and therefore preferred.

While our paper makes an important step toward understanding the efficacy and workings of collective punishment, we believe there are several important dimensions which go beyond its scope. Without the ability to commit to collective punishment participants may not hold consistent or sufficiently strong expectations that the group will be punished in case of low total contributions. One may consequently wonder whether collective punishment without commitment could be effective if the fraction of those willing to engage in it is high enough, so to create the expectation that it will follow under low total contributions. Experimentally, this could be achieved, e.g., by manipulating its cost or by choosing a leader who is in charge of collective punishment for a prolonged period of time.

In the present paper, the collective punishment decision is taken by members of the group. This seems to be a realistic description of many political institutions or certain economic organizational structures such as workers' cooperatives or self-managing work teams but is not an accurate description of other organizations which feature a hierarchy between a principal and a group of agents. Hence, it may be interesting to studying the interplay between a principal in charge of collective sanctions and the internal and external dynamics of a group of agents at the receiving end. It seems to be natural in such a setting that, while the principal holds the power, individual agents have superior information about each others' conduct. Possible areas

of interest include information sharing of agents with the principal and peer-to-peer punishment and ostracism among agents.<sup>22</sup>

Further, even without ensured collective sanctioning, allowing participants to opt into collective punishment, instead of exogenously enforcing it, may already help to maintain high and stable contribution rates. Nalbantian and Schotter (1997) who study forcing contracts, which correspond to exogenously imposed collective punishment, find contribution levels were decreasing over time while we see an increase in our experiment giving individuals to choose to be sanctioned collectively. Opting into punishment institutions provides players with means to signal their willingness to punish, thus amplifying its deterrent effect. Kosfeld *et al.* (2009) and Markussen *et al.* (2014) document an efficiency enhancing effects of institution choice, and find that subjects prefer to form such institutions. Similarly, the literature on peer-to-peer punishment also finds that allowing people to choose increases cooperation rates (see, e.g., Sutter *et al.*, 2010, or Mellizo *et al.*, 2017).

One may also wonder about situations where collective punishment and individual punishment, either peer-to-peer or applied by a central authority, exist in parallel. Dickson (2007) analyzes in detail the interplay between an outside authority who can exert collective punishment and a group of players who can additionally engage in peer-to-peer punishment. This work differs from the literature on standard peer-to-peer punishment in many aspects other than the presence of collective punishment. In particular, private punishment by group members is more expensive compared to most literature on standard punishment and the amount of maximally allowed punishment is fairly restrictive (well below the level required to impede free riding with selfish preferences). Most importantly, a comparison of a setting with collective punishment to one without collective punishment is missing which makes it impossible to assess the efficacy of collective punishment in the Dickson (2007) setting.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/psrm.2023.52>. To obtain replication material for this article, <https://doi.org/10.7910/DVN/6ZGU1X>.

**Acknowledgements.** We thank Eric Dickson as well as seminar audiences at the Universities of East Anglia, Fribourg, Heidelberg, Linz, Lund, Siena, St. Andrews, and Venice for helpful comments and suggestions. We are indebted to Sara Godoy for most valuable assistance in running the experiments as well as to Sandra Miltenyte and Axel Skantze for excellent research assistance. Erik Mohlin is grateful to Handelsbankens forskningsstiftelser (grant #P2016-0079:1), the Swedish Research Council (grant #2015-01751), the Oxford Economic Papers Fund, and the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellowship 2016-0156) for their financial support. Weidenholzer acknowledges support through a BA/Leverhulme Small Grant.

## References

- Agrawal A** (2005) Environmentalism: community, intimate government, and the making of environmental subjects in Kumaon, India. *Current Anthropology* **46**, 161–190.
- Allen SH and Lektzian DJ** (2013) Economic sanctions: a blunt instrument?. *Journal of Peace Research* **50**, 121–135.
- Alventosa A, Antonioni A and Hernández P** (2021) Pool punishment in public goods games: how do sanctioners' incentives affect us?. *Journal of Economic Behavior & Organization* **185**, 513–537.
- Ambrus A and Greiner B** (2012) Imperfect public monitoring with costly punishment: an experimental study. *American Economic Review* **102**, 3317–3332.
- Balafoutas L and Nikiforakis N** (2012) Norm enforcement in the city: a natural field experiment. *European Economic Review* **56**, 1773–1785.
- Baldassarri D and Grossman G** (2011) Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences* **108**, 11023–11027.
- Baldwin K** (2013) Why vote with the chief? Political connections and public goods provision in Zambia. *American Journal of Political Science* **57**, 794–809.
- Baldwin K** (2019) Elected maps, traditional chiefs, and local public goods: evidence on the role of leaders in co-production from rural Zambia. *Comparative Political Studies* **52**, 1925–1956.

<sup>22</sup>In the absence of collective punishment, Carpenter *et al.* (2017) observe in a production setting that profit sharing among agents entices them to report shirking of other agents and in turn leads to increased effort provision.



- Baldwin K and Raffler P.** (2019) Traditional leaders, service delivery, and electoral accountability. In Rodden JA and Wibbels E (eds). *Decentralized governance and accountability: Academic research and the future of donor programming*. Cambridge: Cambridge University Press, pp. 61–90.
- Balliet D** (2010) Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution* **54**, 39–57.
- Battalio R, Samuelson L and Van Huyck J** (2001) Optimization incentives and coordination failure in laboratory stag hunt games. *Econometrica* **69**, 749–764.
- Blackstone W** (1966) *Commentaries on the Laws of England: 1765–1769*, vol. 27. New York: Dawson.
- Bornstein G and Weisel O** (2010) Punishment, cooperation, and cheater detection in “noisy” social exchange. *Games* **1**, 18–33.
- Bowler S and Donovan T** (2002) Democracy, institutions and attitudes about citizen influence on government. *British Journal of Political Science* **32**, 371–390.
- Boyd R and Richerson PJ** (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* **13**, 171–195.
- Boyd R, Gintis H and Bowles S** (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328**, 617–620.
- Brandts J and Cooper DJ** (2006) A change would do you good... an experimental study on how to overcome coordination failure in organizations. *American Economic Review* **96**, 669–693.
- Bruun N and Johansson C** (2014) Sanctions for unlawful collective action in the Nordic countries and Germany. *International Journal of Comparative Labour Law and Industrial Relations* **30**, 253–271.
- Bueno de Mesquita E and Dickson ES** (2007) The propaganda of the deed: terrorism, counterterrorism, and mobilization. *American Journal of Political Science* **51**, 364–381.
- Butler DM and Kousser T** (2015) How do public goods providers play public goods games?. *Legislative Studies Quarterly* **40**, 211–240.
- Camerer CF** (2003) *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.
- Carpenter J, Robbett A and Akbar PA** (2017) Profit sharing and peer reporting. *Management Science* **64**, 4261–4276.
- Castillo JG and Hamman J** (2021) Political accountability and democratic institutions. An experimental assessment. *Journal of Experimental Political Science* **8**, 128–144.
- Centola D, Willer R and Macy M** (2005) The emperor’s dilemma: a computational model of self-enforcing norms. *American Journal of Sociology* **110**, 1009–1040.
- Chang HI, Dawes CT and Johnson T** (2018) Political inequality, centralized sanctioning institutions, and the maintenance of public goods. *Bulletin of Economic Research* **70**, 251–268.
- Chapkovski P** (2021) Strike one hundred to educate one: measuring the efficacy of collective sanctions experimentally. *PLoS ONE* **16**, e0248599.
- Cox M, Arnold G and Tomás SV** (2010) A review of design principles for community-based natural resource management. *Ecology and Society* **15**, 38 [online].
- DeCaro DA, Janssen MA and Lee A** (2015) Synergistic effects of voting and enforcement on internalized motivation to cooperate in a resource dilemma. *Judgment and Decision Making* **10**, 511–537.
- Dickson ES** (2007) On the (in) effectiveness of collective punishment: an experimental investigation. Technical report, New York University.
- Dickson ES, Gordon SC and Huber GA** (2009) Enforcement and compliance in an uncertain world: an experimental investigation. *The Journal of Politics* **71**, 1357–1378.
- Dickson ES, Gordon SC and Huber GA** (2015) Institutional sources of legitimate authority: an experimental investigation. *American Journal of Political Science* **59**, 109–127.
- Dreber A, Rand DG, Fudenberg D and Nowak MA** (2008) Winners don’t punish. *Nature* **452**, 348–351.
- Drumbl MA** (2004) Collective violence and individual punishment: the criminality of mass atrocity. *Northwestern University Law Review* **99**, 539.
- Eckel CC, Fatas E and Wilson R** (2010) Cooperation and status in organizations. *Journal of Public Economic Theory* **12**, 737–762.
- Falletti TG and Riofrancos TN** (2018) Endogenous participation: strengthening prior consultation in extractive economies. *World Politics* **70**, 86–121.
- Fehr E and Gächter S** (2000) Cooperation and punishment in public goods experiments. *American Economic Review* **90**, 980–994.
- Fehr E and Gächter S** (2002) Altruistic punishment in humans. *Nature* **415**, 137–140.
- Feri F, Irlenbusch B and Sutter M** (2010) Efficiency gains from team-based coordination—large scale experimental evidence. *American Economic Review* **100**, 1892–1912.
- Finkel SE** (1987) The effects of participation on political efficacy and political support: evidence from a West German Panel. *The Journal of Politics* **49**, 441–464.
- Fung A and Wright EO** (2001) Deepening democracy: innovations in empowered participatory governance. *Politics & Society* **29**, 5–41.

- Gächter S, Renner E and Sefton M (2008) The long-run benefits of punishment. *Science* **322**, 1510–1510.
- Gelman A, Fagan J and Kiss A (2007) An analysis of the New York City police department’s “stop-and-frisk” policy in the context of claims of racial bias. *Journal of the American Statistical Association* **102**, 813–823.
- Gintis H, Bowles S, Boyd RT and Fehr E (2005) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, vol. 6. Cambridge: MIT Press.
- Goldfrank B (2012) The World Bank and the globalization of participatory budgeting. *Journal of Deliberative Democracy* **8**.
- Gordon J (1999) A peaceful, silent, deadly remedy: the ethics of economic sanctions. *Ethics & International Affairs* **13**, 123–142.
- Grechenig K, Nicklisch A and Thöni C (2010) Punishment despite reasonable doubt—a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies* **7**, 847–867.
- Grossman G and Baldassarri D (2012) The impact of elections on cooperation: evidence from a lab-in-the-field experiment in Uganda. *American Journal of Political Science* **56**, 964–985.
- Hamman JR, Weber RA and Woon J (2011) An experimental investigation of electoral delegation and the provision of public goods. *American Journal of Political Science* **55**, 738–752.
- Hayashi N, Ostrom E, Walker J and Yamagishi T (1999) Reciprocity, trust, and the sense of control: a cross-societal study. *Rationality and Society* **11**, 27–46.
- Heap SPH, Tsutsui K and Zizzo DJ (2020) Vote and voice: an experiment on the effects of inclusive governance rules. *Social Choice and Welfare* **54**, 111–139.
- Heckathorn DD (1988) Collective sanctions and the creation of prisoner’s dilemma norms. *American Journal of Sociology* **94**, 535–562.
- Heckathorn DD (1989) Collective action and the second-order free-rider problem. *Rationality and Society* **1**, 78–100.
- Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E and Henrich N (2010) Markets, religion, community size, and the evolution of fairness and punishment. *Science* **327**, 1480–1484.
- Herrmann B, Thöni C and Gächter S (2008) Antisocial punishment across societies. *Science* **319**, 1362–1367.
- Hilbe C, Traulsen A, Röhl T and Milinski M (2014) Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. *Proceedings of the National Academy of Sciences* **111**, 752–756.
- Holmström B (1982) Moral hazard in teams. *The Bell Journal of Economics* **13**, 324–340.
- Hong S and Cho BS (2018) Citizen participation and the redistribution of public goods. *Public Administration* **96**, 481–496.
- Kosfeld M, Okada A and Riedl A (2009) Institution formation in public goods games. *American Economic Review* **99**, 1335–1355.
- Ledford JGE, Lawler IEE and Mohrman SA (1995) Reward innovations in Fortune 1000 companies. *Compensation & Benefits Review* **27**, 76–80.
- Levinson DJ (2003) Collective sanctions. *Stanford Law Review* **56**, 345–428.
- Magalhães PC (2016) Economic evaluations, procedural fairness, and satisfaction with democracy. *Political Research Quarterly* **69**, 522–534.
- Mansuri G and Rao V (2004) Community-based and-driven development: a critical review. *The World Bank Research Observer* **19**, 1–39.
- Mansuri G and Rao V (2012) Localizing development: does participation work?.
- Markussen T, Putterman L and Tyran J-R (2014) Self-organization for collective action: an experimental study of voting on sanction regimes. *Review of Economic Studies* **81**, 301–324.
- McGillivray F and Smith A (2000) Trust and cooperation through agent-specific punishments. *International Organization* **54**, 809–824.
- McGillivray F and Smith A (2006) Credibility in compliance and punishment: leader specific punishments and credibility. *The Journal of Politics* **68**, 248–258.
- Mellizo P, Carpenter J and Matthews PH (2017) Ceding control: an experimental analysis of participatory management. *Journal of the Economic Science Association* **3**, 62–74.
- Miguel E and Gugerty MK (2005) Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of Public Economics* **89**, 2325–2368.
- Morton RB and Ou K (2015) What motivates bandwagon voting behavior: altruism or a desire to win?. *European Journal of Political Economy* **40**, 224–241.
- Morton RB and Ou K (2019) Public voting and prosocial behavior. *Journal of Experimental Political Science* **6**, 141–158.
- Nalbantian HR and Schotter A (1997) Productivity under group incentives: an experimental study. *The American Economic Review* **87**, 314–341.
- Ockenfels A, Sliwka D and Werner P (2014) Bonus payments and reference point violations. *Management Science* **61**, 1496–1513.
- O’Gorman R, Henrich J and Van Vugt M (2009) Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences* **276**, 323–329.
- Oliver P (1980) Rewards and punishments as selective incentives for collective action: theoretical investigations. *American Journal of Sociology* **85**, 1356–1375.

- Ostrom E** (1999) Coping with tragedies of the commons. *Annual Review of Political Science* **2**, 493–535.
- Ostrom E, Walker J and Gardner R** (1992) Covenants with and without a sword: self-governance is possible. *American Political Science Review* **86**, 404–417.
- Powell GB and Powell GB Jr** (2000) *Elections as Instruments of Democracy: Majoritarian and Proportional Visions*. New Haven: Yale University Press.
- Riedl A, Rohde IM and Strobel M** (2015) Efficient coordination in weakest-link games. *The Review of Economic Studies* **83**, 737–767.
- Scholz JT and Lubell M** (1998) Trust and taxpaying. Testing the heuristic approach to collective action. *American Journal of Political Science* **42**, 398–417.
- Scholz JT and Pinney N** (1995) Duty, fear, and tax compliance: The heuristic basis of citizenship behavior. *American Journal of Political Science* **39**, 490–512.
- Sigmund K** (2007) Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology & Evolution* **22**, 593–600.
- Soss J and Weaver V** (2017) Police are our government: politics, political science, and the policing of race–class subjugated communities. *Annual Review of Political Science* **20**, 565–591.
- Sunstein CR and Ullmann-Margalit E** (1999) Second-order decisions. *Ethics* **110**, 5–31.
- Sutter M, Haigner S and Kocher MG** (2010) Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies* **77**, 1540–1566.
- Tavits M** (2007) Clarity of responsibility and corruption. *American Journal of Political Science* **51**, 218–229.
- Taylor M** (1982) *Community, Anarchy and Liberty*. Cambridge: Cambridge University Press.
- Touchton M and Wampler B** (2014) Improving social well-being through new democratic institutions. *Comparative Political Studies* **47**, 1442–1469.
- Van Der Merwe C** (2015) *European Condominium Law*. Cambridge: Cambridge University Press.
- Van Huyck JB, Battalio RC and Beil RO** (1990) Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review* **80**, 234–248.
- Wagner AF, Schneider F and Halla M** (2009) The quality of institutions and satisfaction with democracy in Western Europe —a panel analysis. *European Journal of Political Economy* **25**, 30–41.
- Warren ME** (2011) Voting with your feet: exit-based empowerment in democratic theory. *American Political Science Review* **105**, 683–701.
- Weber RA** (2006) Managing growth to achieve efficient coordination in large groups. *American Economic Review* **96**, 114–126.
- Wheelock DC and Wilson PW** (2013) The evolution of cost-productivity and efficiency among us credit unions. *Journal of Banking & Finance* **37**, 75–88.
- Willer R, Kuwabara K and Macy MW** (2009) The false enforcement of unpopular norms. *American Journal of Sociology* **115**, 451–490.
- Yamagishi T** (1986) The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* **51**, 110.