Mengel, Friederike; Mohlin, Erik; Weidenholzer, Simon

Working Paper

Collective Incentives and Cooperation in Teams with Imperfect Monitoring

Working Paper, No. 2018:11

Provided in Cooperation with:
Department of Economics, School of Economics and Management, Lund University

Suggested Citation: Mengel, Friederike; Mohlin, Erik; Weidenholzer, Simon (2018) : Collective Incentives and Cooperation in Teams with Imperfect Monitoring, Working Paper, No. 2018:11, Lund University, School of Economics and Management, Department of Economics, Lund

This Version is available at:
https://hdl.handle.net/10419/260240

Working Paper 2018:11

Department of Economics
School of Economics and Management

Collective Incentives and Cooperation in Teams with Imperfect Monitoring

Friederike Mengel
Erik Mohlin
Simon Weidenholzer

May 2018

LUND UNIVERSITY

Collective Incentives and Cooperation in Teams with Imperfect Monitoring *

Friederike Mengel † Erik Mohlin ‡ Simon Weidenholzer§

May 27, 2018

Abstract

We experimentally explore the role of collective incentives in sustaining cooperation in finitely repeated public goods games with imperfect monitoring. In our experiment players only observe noisy signals about individual contributions, while total output is perfectly observed. We consider sanctioning mechanisms that allow agents to commit to collective punishment in case total output fall short of a target. We find that cooperation is higher in the case of collective punishment compared to both the case of no punishment and the case of standard peer-to-peer punishment which conditions on the noisy signals. Further experiments indicate that both the commitment possibility and the collective nature of punishment matter for the positive effect of collective incentives on cooperation.

*We thank seminar audiences at the Universities of East Anglia, Fribourg, Heidelberg, Linz, Lund, Siena, 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), and the Oxford Economic Papers Fund for their financial support. Weidenholzer acknowledges support through the BA/Leverhulme Small Research Grants scheme.

†*Affiliation*: University of Essex and Lund University *Address*: Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. *E-mail*: fr.mengel@gmail.com

‡*Affiliation*: Lund University. *Address*: Department of Economics, Lund University, Sweden. Address: Tycho Brahes väg 1, 220 07 Lund, Sweden. *E-mail*: erik.mohlin@nek.lu.se.

§*Affiliation*: University of Essex. *Address*: Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. *E-mail*: sweide@essex.ac.uk

1 Introduction

Many social interactions feature a tension between individual incentives and group benefits. Examples include effort provision in teams, voluntary public goods provision, common pool resources, collusion in oligopolies, and arms races. The underlying conflict is perhaps made most explicit in (finitely repeated) public goods games. In such games self-interested individuals prefer to free ride and enjoy the benefits of the public good without contributing themselves. At the same time the payoff of the group would be maximized if everybody contributed. Consequently the literature has turned to the question how individuals could be stimulated to contribute more. It is well-documented in an experimental context that the ability to punish non-contributors significantly increases contributions and allows societies to move towards socially optimal levels of contributions (see e.g. Ostrom, Walker, and Gardner 1992 and Fehr and Gächter 2000, 2002). Despite costly punishment being at odds with selfish preferences, a considerable fraction of individuals is willing to sacrifice own payoffs to punish non-cooperators, which increases cooperation. 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 literature (see e.g. Bornstein and Weisel 2010, Grechenig, Nicklisch, and Thöni 2010, Ambrus and Greiner 2012).

In this paper we study under which conditions, if any, groups can maintain high contribution and welfare levels, when monitoring is imperfect. In particular, we focus on collective sanctioning mechanisms where all members of a group are subject to punishment, regardless of their individual contributions (or the noisy signals thereof). Instead of individual contributions, sanctions are anchored on group outcomes which are typically observed in a much less noisy way, as stressed by e.g. Holmström (1982).

There are many real life instances where groups are subject to collective sanctions, including punishment in military and educational institutions.[1] Moreover, a variety of salary schemes feature collective elements where remuneration is conditional on group or team performance. For instance, a 1993 survey by Ledford, Lawler, and Mohrman (1995) revealed that 70% of Fortune 1000 companies used some form of team incentives and that 76% of self-managing work teams use team incentives. Collective punishment and team incentives are related in the sense that both condition on the performance of a group.

[1] A notable (and extreme) example is provided by the ancient Roman practice of decimation where one in ten soldiers of a military unit was randomly chosen to be executed. More recent applications of this practice include the execution of 1 in 10 soldiers of a 120 strong Company of the 141st Catanzaro Infantry Brigade, which had mutinied, during WW1 and the black bean lottery during the Texan revolution where 17 Texans, chosen by lot out of 176 escapees, were executed.

The former is framed in terms of punishment and the latter is typically associated with rewards such as bonuses. Since withholding of a bonus can be interpreted as a form of punishment, the two can be viewed as opposite sides of the same coin.[2]

Collective incentives transform an underlying public goods game into a coordination game. In this coordination game selfish individuals 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, the resulting coordination game features an equilibrium where everybody contributes and an equilibrium where nobody contributes.

Often the decision to reward or punish a group of agents is taken by an outside authority or principal. In other cases, notably workers' cooperatives or self-managing work teams, the group members can (at least to a certain degree) decide for themselves whether to impose a sanction or not depending on whether production targets have been reached or not. Holmström (1982) points out that such a scheme involves a commitment problem: once a team has produced a surplus they have no incentive to destroy it or give it away even if it falls below the target. 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 team members can commit in advance to a mechanism that punishes everyone in case production is below a target, then they can choose beforehand whether team production is characterized by a public goods game or by a coordination game. Further, if punishment is committed this may provide a focal point and signal team members to coordinate on the high contribution-high welfare equilibrium of the resulting coordination game.

We experimentally investigate the potential of collective sanctions and commitment to increase contributions and welfare when information about individual contributions is noisy. To this end, we study finitely repeated public goods games where subjects receive noisy signals on the contributions of others but the sum of contributions is observed without noise. Our treatments vary the punishment technology subjects have at their disposal.

In a first step, we compare contribution and welfare levels under (i) the possibility to commit to collective punishment with (ii) standard peer-to-peer punishment (as studied in the previous literature) and (iii) the case where subjects may not punish at all. We find that contribution rates are higher in the presence of the possibility to commit to collective punishment than under either the absence of the possibility to punish or standard peer-to-peer punishment. Welfare comparisons of the different punishment mechanisms depend on whether punishments and punishment costs are assumed to be monetary or not. If these costs are monetary, as in the case of fines or bonuses, they amount to a redistribution

[2]There is also some empirical evidence that receiving a bonus lower than an expected amount decreases employee satisfaction (Ockenfels, Sliwka, and Werner 2014).

of welfare, and consequently the overall welfare effects mirror the effects on contributions. In this paper we follow the majority of the literature and interpret punishments and punishment costs as non-monetary, so that punishment corresponds to a destruction of resources. We find that welfare is significantly higher with the possibility to commit to collective punishment than in either the standard peer-to-peer punishment case or the no-punishment case, though the latter effect is only marginally significant. 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 towards unambiguously positive long-run welfare effects of being able to commit to collective punishment.[3]

Additional evidence for the effectiveness of collective punishment can be obtained by dividing groups into those where collective punishment is in fact committed and those where it is not committed. 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. Strikingly, when punishment is committed, the contribution level is well above 90% in the final round of our experiment. The presence of committed collective punishment thus allows subjects to coordinate on the high contribution equilibrium in the induced coordination game.

Having established the efficacy of collective punishment with commitment, we proceed to assess the relative importance of the two factors *collectiveness* and *commitment* in the punishment technology. To this end, we have run two additional treatments: The first treatment features collective punishment without commitment, i.e. subjects may decide after the contribution stage whether they want to punish everybody in the group. The second treatment augments standard peer-to-peer punishment with commitment, i.e. subjects can commit before the contribution stage to punish others conditional on the noisy signal that will be received. With respect to the first treatment we find that collective punishment without commitment is incapable of raising either contributions or payoff levels. This finding underscores the importance of commitment for collective punishment to work, thus proving support for Holmström's (1982) assertion on moral hazard in teams. We speculate that the fraction of individuals who are willing to engage in costly collective punishment is simply not high enough to create expectations that collective punishment will be implemented with sufficiently high probability.

Analysing data on standard peer-to-peer punishment with commitment, reveals that it may result in (marginally significantly) higher contribution levels than standard peer-to-peer punishment (despite not being part of an equilibrium in the present setting). The

[3]The previous literature 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, Rand, Fudenberg, and Nowak (2008). However, in the long run the frequency of punishment decreases and social welfare with punishment may be higher (Gächter, Renner, and Sefton 2008).

threat of being automatically punished (justified or not) seems to induce individuals to contribute more. Note, however, that in the presence of noise committed standard peer-to-peer punishment is rather wasteful since subjects are wrongfully punished all too often. In fact, roughly twice as much resources are lost under standard peer-to-peer punishment as compared to any other punishment technology. This is also the main reason why the potentially positive effects of increased contributions do not translate into increased welfare.

We further consider the case of strong peer-to-peer punishment which has been suggested by Ambrus and Greiner (2012) to be capable of guaranteeing fairly high contribution and payoff levels in a low noise environment. Under strong peer-to-peer punishment subjects can subtract more punishment points at a lower cost from others. We too find a positive effect of strong punishment on contribution rates, our results on welfare are however strikingly different to those of Ambrus and Greiner (2012). Strong punishment leads to sizeable welfare losses as compared to standard punishment or the no punishment benchmark, with payoff levels being roughly half of those in the latter case. While strong punishment generates fairly high contribution rates, the welfare loss incurred through punishment by far outweighs this positive aspect. We attribute this finding to our environment featuring a substantially higher noise, where contributors are much more frequently labelled as non-contributors. Thus, while strong punishment may work in low noise environments it can be counterproductive when noise levels are higher. Collective punishment with commitment seems to be most effective in such circumstances.

The only paper on collective punishment we are aware of where the punishment decision is endogenous Dickson (2007) who analyses 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 rather expensive (at three times the usual cost). Further, the amount of maximally allowed punishment is fairly restrictive (well below the level required to impede free riding with selfish preferences). Last 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.

Our work complements existing literature along several further dimensions. Closest to the current study are a number of papers studying the implications of imperfect monitoring in public goods games. Ambrus and Greiner (2012) study standard peer-to-peer punishment under imperfect monitoring, where contributors are incorrectly identified as non-contributors with a 10% probability and non-contributors are always detected. They find that even for such low noise levels total earnings are lower with access to punishment

than without. As discussed above, they also consider a stronger form of punishment and find that it enables contributions to be stabilized at a high level, generating total earnings are higher than in the absence of the punishment opportunity, though the frequency of punishment is higher than in the noiseless benchmark. In contrast, we document a welfare diminishing effect of strong punishment in high noise environments. Grechenig, Nicklisch, and Thöni (2010) confirm the welfare diminishing role of imperfect monitoring by showing that contributions are decreasing in the level of noise. Under high noise levels, punishment is not able to raise contributions, let alone social welfare. Ambrus and Greiner (2017) study a form of individual punishment where - unlike in peer-to-peer punishment - group members democratically vote on whom to punish. Still, punishment in their study is individual, not collective. They document that this leads to higher contribution and welfare levels then standard peer-to-peer punishment under both perfect monitoring and imperfect monitoring (in the same information structure as in Ambrus and Greiner (2012)). This result is mainly driven by anti-social punishment being curbed under democratic voting. Similarly, Fischer, Grechenig, and Meier (2016) feature centralized punishment where only one agent may punish others. This too may lead to less anti-social punishment under imperfect monitoring.[4]

Related to imperfect monitoring of actions are settings where there is uncertainty regarding how much individuals can contribute. In Patel, Cartwright, and Van Vugt (2010) only some individuals have the option to contribute in each round and their identity is not known. Under this alternative formulation individuals contribute significantly less than in the noiseless benchmark. Similarly, the punishment frequency is also lower. In Bornstein and Weisel (2010) subjects differ in the endowment they can contribute in every period. It is, thus, no longer clear whether relatively low contributions result from low endowments or are simply low contributions per se. Indeed, the ability to hide one's true contributions (behind random draws of endowments) lowers contributions and punishment frequency as compared to a scenario where endowments are observable.

Another branch of the literature studies incentive structures that are based on team performance. Most prominently collective- or group punishment has been proposed by Holmström (1982) to overcome the moral hazard in teams problem. As mentioned above, Holmström underscores the importance of commitment by pointing out that in its absence team members will not find it optimal to destroy any of the just produced surplus. Thus, an outside principal who punishes the group members by withholding surplus is essential for group punishment to work. Our finding that commitment is needed for collective punishment to be effective corroborates this prediction. Nalbantian and Schotter (1997) experimentally study, amongst other incentive structures, the case of forcing contracts

[4]Imperfect monitoring of others' actions may lead to less cooperation even in the absence of the opportunity to punish, as demonstrated by e.g. Bereby-Meyer and Roth (2006), Fudenberg, Rand, and Dreber (2012), and Abbink and Sadrieh (2009).

where subjects receive a relatively high wage if the group effort reaches a certain level and receive a low wage otherwise. Forcing contracts, thus, correspond to exogenously imposed collective punishment. While forcing contracts are initially capable to raise effort, effort provision is declining over time and by the end of the experiment is no larger than in the no punishment case. In contrast, in our framework where subjects can decide whether to implement collective punishment, contributions stay at fairly high levels throughout the experiment. Moreover, for groups where collective punishment is committed contributions are even increasing over time. Thus, allowing subjects to opt into collective punishment, instead of exogenously enforcing it, may help to maintain high and stable contribution rates.

Our work also speaks to previous work on the choice of punishment institutions under perfect monitoring, since agents are automatically punished in our committed punishment treatments. Opting into punishment institutions provides players with means to commit to punishment, thus amplifying its deterrent effect. Kosfeld, Okada, and Riedl (2009) document clear welfare enhancing effects of institution choice. Most of the time institutions are formed and more people than predicted by theory join such institutions. Markussen, Putterman, and Tyran (2014) confirm the efficiencyiy enhancing effect of institution choice by showing that subjects indeed would vote to form such institutions provided they are relatively cheap. Our results expand the scope of this literature by documenting welfare enhancing effects of (collective) punishment institutions under imperfect monitoring.

Collective punishment gives rise to a weakest link coordination game. [5] As first observed by Van Huyck, Battalio, and Beil (1990), it is rather difficult to achieve coordination on high effort equilibria in such coordination games. A number of contributions have consequently examined factors that may lead to the emergence of high effort equilibria. Examples include: temporally changing incentive structures, (Brandts and Cooper 2006) gradually growing group sizes (Weber 2006), decision making by teams (Feri, Irlenbusch, and Sutter 2010) or network formation (Riedl, Rohde, and Strobel 2015). 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. More specifically, our players can choose between a public goods game and a coordination game. Choosing the latter may provide a focal point for the high effort equilibrium in this game.

The rest of the paper is organized as follows. Section 2 gives some theoretical background and introduces the different punishment regimes. Section 3 presents our experimental design. Our results are reported in Section 4. Section 5 concludes.

[5] Admittedly, in contrast to much of the literature our setting features only two possible effort levels.

2 Public Good Games with Punishment and Imperfect Monitoring

In this section we introduce the public good game and the different punishment institutions and derive theoretical predictions. 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, $g_i \in \{0, e\}$.[6] 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 $j \neq i$. This signal reveals the true contribution level with a signal accuracy of $\Pr(s_{ij} = g | g_j = g) \in (0.5, 1]$. 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

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

where α, 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. On the other hand, welfare maximization dictates contribution levels of e for all participants, resulting in payoffs of $\alpha n e$.

We are interested in the circumstances under which subjects can be enticed to contribute to the public good and consequentially achieve high levels of welfare. To this end, we discuss and analyse several punishment regimes.

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.[7] 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 βP for some $\beta \in \mathbb{R}$, where typically $\beta < 1$. We assume that all members of a group share the cost of punishment equally, so that in case of collective punishment each group members' payoff is lowered by $P(1 + \beta)$. We focus on two different scenarios, depending on whether the decision 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 or not.

[6] Binary contributions allow for a fairly straight forward and for subjects comprehensible implementation of imperfect monitoring.

[7] 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.

2.1.1 Collective Punishment without Commitment

Suppose the decision whether to collectively punish or not is taken after the public goods game has been played. One agent is drawn at random and decides whether to punish the group or not.[8,9] 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 subgame perfect equilibrium no player will contribute.

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 subjects have committed 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 or not. 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.[10]

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. The presence of collective punishment gives rise to a strategic structure akin to weakest link coordination games as studied in Van Huyck, Battalio, and Beil (1990). Thus, the profile where everybody contributes and the profile where nobody contributes are both Nash equilibria. The payoffs

[8] It is a straightforward exercise to generalize this to a large class of collective decision making mechanisms such as majority vote.

[9] That is in the present setting players can either punish or not. This formulation allows us to present various forms of punishment with commitment, studied later on, in the simplest way possible to subjects.

[10] 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.

in these states are given by αne and $e - p(1+\beta)$, respectively. Thus, the socially optimal payoff is achieved in one of this subgame's equilibria.

We now turn to the question whether our randomly selected player j will implement collective punishment or not. Given the equilibria in the subgames there are two subgame 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). Note that there does not exist a subgame 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

In addition to the collective punishment regimes discussed above we are interested in peer-to-peer punishment. Again we distinguish between whether punishment can be committed in advance or not.

2.2.1 Peer-to-Peer Punishment without Commitment

Under the most commonly studied punishment regime the public good game is augmented with a peer-to-peer punishment technology, where after the actual public good each player i can decide whether to punish each player j or not, $p_{ij}^S \in \{0, 1\}$. In case of punishment P points will be subtracted from the punished, resulting in a cost of βP to the punisher. Under standard peer-to-peer punishment the payoff function is given by,

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

Note that as punishment is costly no rational and selfish player will engage in it. This observation holds regardless of whether information about contributions is noisy or not. Thus, in the only subgame perfect equilibrium of this game no player will be punished and no player will contribute.

2.2.2 Peer-to-Peer Punishment with Commitment

Our final punishment technology features standard peer-to-peer punishment with commitment. 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 or not. Punishment is contingent on the noisy signal received by player i on j's contributions, s_{ij}. When i commits to 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 standard 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 subgame perfect Nash equilibria where agents commit to punish and 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 exit subgame 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. For the signal precision of 0.6 chosen in our experiment the unique subgame perfect equilibrium is such that nobody contributes and nobody commits to punishment (assuming subjects are risk-neutral selfish money-maximisers)[11]

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 mechanism described in the previous section correspond to a 2×2 factorial design. One dimension is defined by whether there is standard peer-to-peer punishment or collective punishment. The other dimension is defined by whether subjects can commit to punishments or not. In addition we implemented a treatment with a strong punishment technology, see Ambrus and Greiner (2012), which we discuss in Section 4.4.[12] Our main treatments are summarised in Table 1.

In all treatments subjects participated in a 50-period repeated public good game. Subjects were randomly matched in groups of four which remained constant for the entire duration of the experiment. In each period of the public goods game each subject 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

[11] See Appendix A for a formal proof of this statement.

[12] In addition to the treatments featured here we ran treatments with lower noise rates and slightly different information structures. See Appendix D.

Table 1: Main Treatments.

	No commitment	Commitment
Standard Punishment	**S** (2400,12)	**S-Comm** (4600,23)
Collective Punishment	**C** (2600,13)	**C-Comm** (4600,23)
No Punishment	**N** (1600,8)	

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

of possibly excessive punishment, and to make sure that subjects did not leave the 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.[13] 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 $e = 20$.

In addition, 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 behaviour of a third player. With probability 0.6 the signal reflected the true behaviour of a subject. With the remaining probability 0.4 a contributor was labelled as a non-contributor and a non-contributor was labelled as a contributor. In addition, subjects 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 less noisy way than the results of a collaborative effort.

In the no punishment treatment **N** subjects simply played the contribution game without a punishment stage. In the punishment treatments without commitment subjects could decide on punishing other subjects after the contribution stage. In the standard punishment treatment **S** each subject i was asked whether she wanted to subtract $p = 15$ punishment points from each other subject j or not. As usual in the literature, we set the punishment cost 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 subjects. The strong punishment treatment **S-Strong** was exactly the same as the punishment treatment with the exception that now 5 tokens bought 30 punishment points, implying $\beta = \frac{1}{6}$ as in the strong punishment treatments of Ambrus and Greiner (2012).

[13]This limited liability constraint was not reached a single time in our experiment.

In the collective punishment treatment **C** each subject was asked, after receiving 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. Subjects were made aware that the decision of one of the players would be chosen at random for implementation.[14] The fact that everybody had to bear the cost of punishment allows us to directly compare collective to standard punishment. An alternative interpretation is that an individual that collectively punishes simply destroys part of the surplus generated at the contribution stage.

In the punishment treatments with commitment subjects could commit to punishing other subjects 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 standard punishment with commitment treatment **S-Comm** each subject was asked whether they commit to punish another subject if they receive the noisy signal that (i) the subject contributed or (ii) did not contribute. Thus, each subject had to make six decisions at this stage.

In the collective punishment with commitment treatment **C-Comm** subjects were asked whether they commit to punishing everybody (including themselves) in case somebody did not contribute. Before the contribution stage the choice of one subject was randomly implemented and everybody was made aware whether a collective punishment mechanism was in place or not.[15]

The experimental sessions were conducted at ESSEXLab at the University of Essex. Ethical approval was obtained by the Essex Social Sciences Faculty Ethics Committee under Annex B. Participants were recruited using Orsee (Greiner 2004). Subjects received written instructions and were allowed to ask questions before the experiment which were answered in private. The experiments were programmed using zTree (Fischbacher 2007). Overall, 340 participants participated in our main treatments.[16] Sessions lasted approximately 1 hour 45 min, including instructions, a short post experimental questionnaire and payment of subjects. On average subjects were paid approximately GBP 18, including a show up fee of 2.5 GBP.

[14]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 subjects as compared to one constant decision maker, iii) avoids the tie splitting problem with four subjects and iv) avoids certain indifference cases as compared to majority vote.

[15]Note that collective punishment in fact only conditions on whether the sum of contributions is maximal and thus requires even less information than is available here. Admittedly, 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.

[16]This includes the strong punishment treatment reported on in Section 4.4

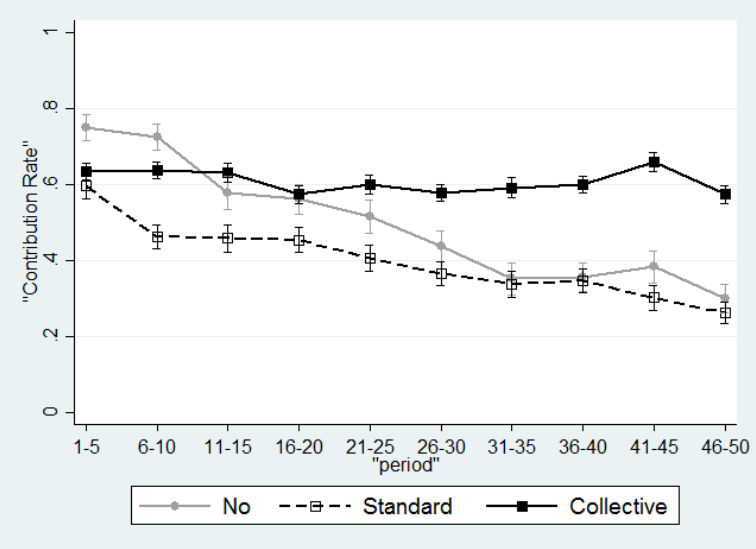
4 Results

4.1 Efficacy of Collective Punishment with Commitment

In a first step we aim to assess the efficacy of collective punishment with commitment. We focus on the collective punishment treatment *with* commitment as a benchmark, since this seems the natural setting for collective punishment schemes and the way most of these schemes operate in practice. We hence 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), as featured in most of the literature on cooperation and punishment in public goods, and the case where the possibility of punishment is absent (treatment N). Later, in Section 4.3 we will try to gauge the relative importance of the effects of commitment and "collectiveness".

First, we consider contribution rates. Figure 1 plots these 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

Figure 1: Contribution rates across treatments N, S and C-COMM.



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 δ_S and δ_{C-Comm} are dummy variables for treatments S and C-Comm respectively. The constant α captures the baseline effect in the case of no

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 (β_1)	-0.095	-0.202**	-0.154*	-0.036	-0.053
	(0.087)	(0.084)	(0.086)	(0.094)	(0.105)
C-Comm (β_2)	0.107	-0.141*	-0.023	0.238***	0.279***
	(0.064)	(0.074)	(0.068)	(0.068)	(0.077)
period (β_3)		-0.010***			
		(0.001)			
period × **S** (β_4)		0.004**			
		(0.001)			
period × **C-Comm** (β_5)		0.009***			
		(0.001)			
Constant (α)	0.500***	0.761***	0.635***	0.365***	0.341***
	(0.045)	(0.060)	(0.054)	(0.041)	(0.052)
p-value Test $\beta_1 = \beta_2$	0.025**		0.1074	0.009***	0.003***
Observations	8,600	8,600	4,300	4,300	1,720
R-squared	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

punishment. 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. Since we do not expect equilibrium predictions to matter until subjects have had time to learn and adapt their behaviour much of our analysis focuses on the last 10 periods, presented in column (5).

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%-level. In contrast, standard 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 standard 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 standard punishment ($\beta_3+\beta_4 \approx -0.06, p = 0.0001$) while the time trend for collective punishment with commitment is not significantly different from zero ($\beta_3 + \beta_5 \approx 0.00$, p=0.6947). Summarising our findings so far we have:

Result 1. *Contribution levels are significantly higher under collective punishment with commitment compared to the no punishment scenario or to standard punishment. There is no significant difference between the no punishment scenario and standard punishment. While contribution levels are stable under collective punishment with commitment they are decreasing under standard punishment and in the no punishment scenario.*

Collective punishment may thus support 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 2.

Figure 2: Profits

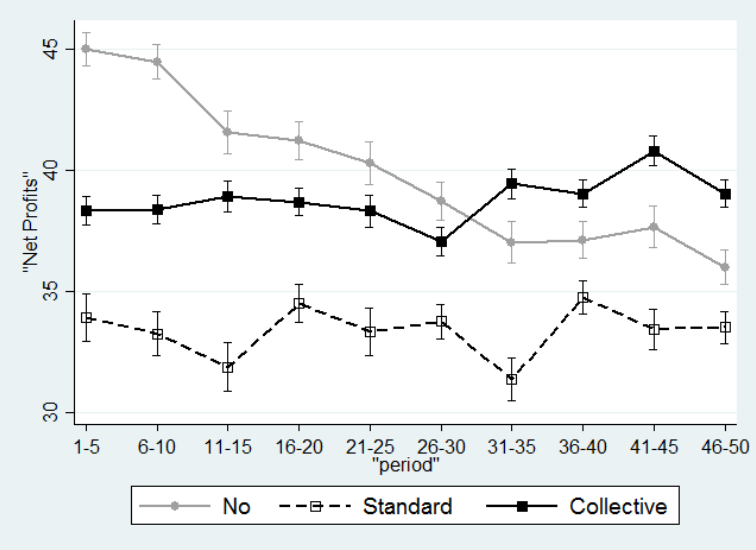


Table 3 reports the results of the following regression

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

where π_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.

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 (β_1)	-6.508***	-11.790***	-9.117***	-3.900*	-3.354
	(1.979)	(2.352)	(2.127)	(2.099)	(2.350)
C-Comm (β_2)	-1.209	-7.271***	-4.257**	1.839	3.231*
	(1.567)	(1.843)	(1.650)	(1.665)	(1.856)
period (β_3)		-0.205***			
		(0.026)			
period × **S** (β_4)		0.207***			
		(0.057)			
period × **C-Comm** (β_5)		0.238***			
		(0.041)			
Constant (α)	40***	45.22***	42.70***	37.30***	36.81***
	(0.906)	(1.200)	(1.082)	(0.828)	(1.043)
p-value Test $\beta_1 = \beta_2$	0.0192**		0.0337**	0.0218**	0.0154**
Observations	8,600	8,600	4,300	4,300	1,720
R-squared	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

We find that collective punishment with commitment has a positive effect on payoffs relative to the baseline of no punishment, even though this effect is only significant at the

10%-level. Standard punishment is associated with lower payoffs than the baseline of no punishment, although this effect is not statistically significant across the last ten periods. However collective punishment with commitment yields significantly higher profit than standard punishment without commitment, as revealed by the t-test of equality of coefficients β_1 and β_2. The effect is substantial and corresponds to an increase of 6.585 ECU, or roughly 20%. 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 standard punishment ($\beta_3 + \beta_4 \approx 0.00, p = 0.9654$). While payoffs seem to be increasing under collective punishment, this effect is not significant ($\beta_3 + \beta_5 \approx 0.03, p = 0.3140$), i.e. we cannot reject $\beta_3 + \beta_5 = 0$.

Result 2. *Payoff levels under collective punishment with commitment are (marginally) significantly higher than in the no punishment scenario and are significantly higher than under standard punishment. Payoff levels in the no punishment case are not significantly higher than under standard punishment. While payoff levels are decreasing in the no punishment scenario they are stable under collective punishment with commitment and under standard punishment without commitment.*

Thus, collective punishment may restore relatively high levels of welfare in a noisy environment where standard punishment does not work as well as in the noiseless case (see also Ambrus and Greiner (2012) or Fischer, Grechenig, and Meier (2016)). Moreover, while positive payoff implications of collective punishment with commitment (as compared to the no-punishment case) start to become evident only towards the end of the experiment, the pattern of time trends documented in column (2) of Tables 2 and 3 points towards long term positive welfare effects. This picture is reinforced by focusing on groups where collective punishment is in fact committed, as we show in the next subsection.

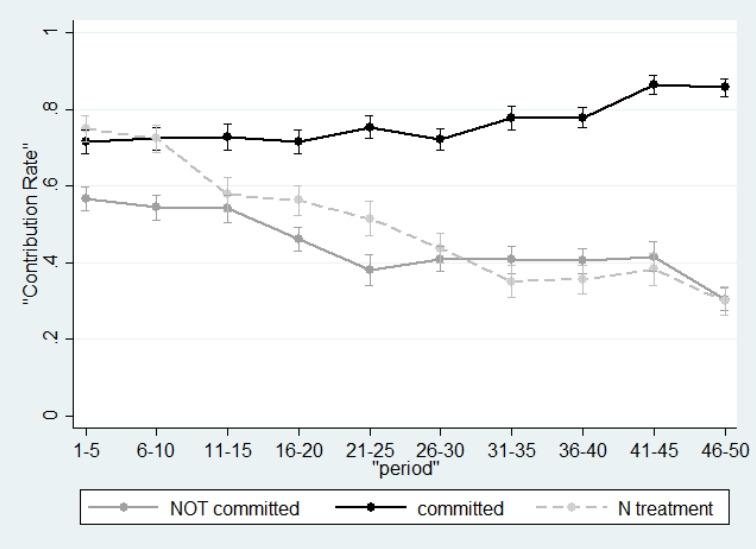
4.2 Collective Punishment as a Coordination Device

As previously noted, whenever collective punishment is committed subjects 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 subgame corresponds to the original public goods game. By choosing whether to implement collective punishment or not 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 subgame 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 however no subgame perfect equilibrium where collective punishment is implemented and the no contribution

equilibrium of the coordination game is played. The presence of collective punishment may, thus, act as a focal point in the implied coordination game.

In order to asses this prediction, we now contrast contribution rates in groups where collective punishment is committed to those where it is not committed. Figure 3 plots contribution rates in these two cases over the course of the experiment. As a benchmark we also include the no punishment (**N**) treatment in the figure.

Figure 3: Contributions with and without collective punishment committed



While behaviour in groups in the collective punishment treatment where collective punishment was not implemented closely resembles behaviour 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% in the last round.

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 don't as different. 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 (see Dal Bo, Foster, and Putterman (2010) for a similar strategy).

Table 4 (upper half) shows the results. Horizontal comparisons illustrate the selection problem. Irrespective of whether collective punishment is implemented or not, there is

Table 4: Contributions and Profits when collective punishment is and is not committed

Votes in Favour	0	1	2	3	4	Average
			Contribution Rates			
Coll. Pun. Implemented	-	0.55	0.76	0.92	0.98	0.81
Coll. Pun. NOT Implemented	0.33	0.44	0.43	0.56	-	0.43
			Profits			
Coll. Pun. Implemented	-	33.8	35.1	44.4	48.1	40.5
Coll. Pun. NOT Implemented	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 favour of implementing the collective punishment mechanism and whether or not it was actually implemented.

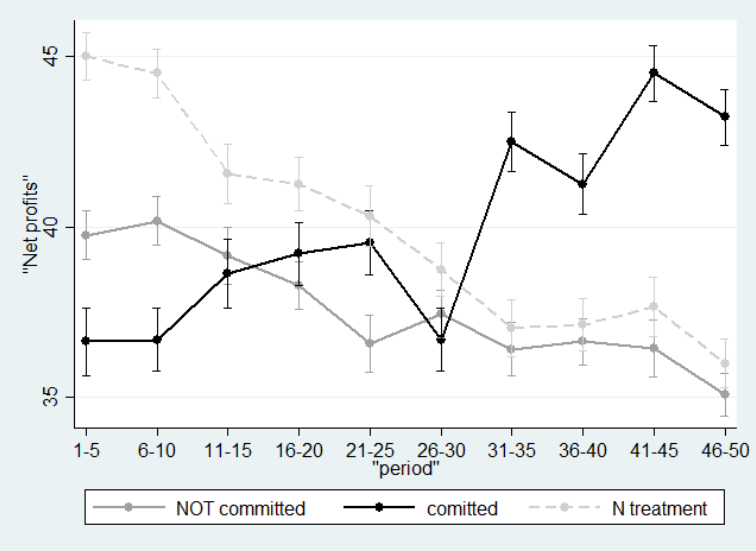
always more cooperation the more people voted in favour of the institution. Presumably this is, because more cooperatively inclined individuals are more likely to vote in favour of establishing the institution. Vertical comparisons illustrate the effect of the institution being implemented. Conditional on the number of people who voted in favour of the institution (1,2 or 3), whether or not 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 favour (t-test, $p = 0.316$). It is highly statistically significant for 2 votes ($p < 0.0001$) and 3 votes ($p = 0.015$).

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.*

It is interesting to contrast this finding to Nalbantian and Schotter (1997) who study forcing contracts, which correspond to exogenously imposed collective punishment, where contribution levels were decreasing over time. The comparison suggests that allowing subjects to opt into collective punishment institutions, instead of exogenously imposing them, may help to maintain high and stable contribution rates. This finding also echoes a number of results from the literature on peer-to-peer punishment who found that allowing people to choose increases cooperation rates (see e.g. Sutter, Haigner, and Kocher (2010), or Mellizo, Carpenter, and Matthews (2017)).

We now proceed to assess the implications of committed collective punishments for welfare. Figure 4 plots the evolution of payoffs in the two subgroups and contrasts it to the no-punishment benchmark. While profits of subjects in groups where collective punishment is not implemented are statistically indistinguishable from profits made by subjects in the no punishment benchmark, the profit of subjects in groups with implemented collective punishment is clearly higher in the later periods of the experiment. Further, payoffs are evidently increasing for subjects in groups where punishment is committed ($\beta = 0.1635^{***}$) and are decreasing for the other two reference groups (non-committed:

Figure 4: Profits with and without collective punishment 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 favour of establishing the institution. Interestingly, in relatively uncooperative groups where only one person voted in favour of the mechanism, payoffs are lower when the mechanism is implemented (though not statistically different, $p = 0.421$). 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 favour ($p = 0.055$) and when three voters are in favour ($p = 0.120$).

Result 4. *On average payoff 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.*

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

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

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 (β_1)	0.179**	0.105	0.138*	0.220**	0.210*
	(0.0858)	(0.078)	(0.0787)	(0.0998)	(0.110)
coll (β_2)	0.0204	0.061	0.0307	0.0101	0.0106
	(0.0889)	(0.079)	(0.0830)	(0.101)	(0.109)
coll×commit (β_3)	0.0254	-0.120	-0.0374	0.0882	0.156
	(0.116)	(0.119)	(0.111)	(0.133)	(0.146)
period		-0.006***			
		(0.001)			
period × commit		0.002			
		(0.002)			
period × coll		-0.001			
		0.002			
period × comm × coll		0.005*			
		(0.003)			
Constant (α)	0.405***	0.558***	0.481***	0.328***	0.287***
	(0.0750)	(0.060)	(0.0679)	(0.0852)	(0.0906)
Observations	9800	9800	4,900	4,900	1,960
p-value $\beta_1 + \beta_3$	0.5427	0.6736	0.9291	0.2592	0.0920*
p-value $\beta_2 + \beta_3$	0.0120**	0.6089	0.2073	0.0010***	0.0004***
R-squared	0.038	0.023	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

Here δ_{Comm} is a dummy indicating that the treatment was one with commitment (S-Comm or C-Comm) and δ_{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 ten periods. The commitment coefficient is significant and positive, indicating that adding commitment to either peer-to-peer punishment or to collective punishment ($\beta_1 + \beta_3$) has a significant positive effect on the rate of cooperation. In contrast, the collectiveness coefficient, though positive, is small and not significant, implying that *without* commitment, collective punishment does not lead to higher contributions. The sum of the collective coefficient and the interaction term coefficient ($\beta_2 + \beta_3$) is, however, marginally significant, showing higher contributions for committed punishment when it is collective.

Result 5. *The ability to commit to punishment increases contributions compared to standard punishment. Whether punishment is collective or not does not significantly influence contributions in the absence of commitment, but it seems to lead to weakly higher contributions with commitment.*

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

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

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	-1.289	-0.358	1.850	1.646
	(2.734)	(3.396)	(2.945)	(2.933)	(3.035)
coll	2.485	5.641**	3.694*	1.277	0.196
	(1.987)	2.390	(2.131)	(2.136)	(2.359)
coll×commit	2.436	-0.613	1.465	3.406	6.159
	(3.399)	(4.25)	(3.646)	(3.635)	(3.818)
period		0.002			
		(0.050)			
period × commit		0.079			
		(0.082)			
period × coll		-0.123**			
		(0.061)			
period × commit × coll		0.119			
		(0.101)			
Constant	33.49***	33.43***	33.58***	33.40***	33.46***
	(1.757)	(2.020)	(1.828)	(1.926)	(2.103)
Observations	9,800	9,800	4,900	4,900	1,960
p-value coll+coll×commit	0.0806*	0.1594	0.0876*	0.1179	0.0395**
p-value comm+coll×commit	0.1216	0.4608	0.6079	0.0181**	0.0015***
R-squared	0.028	0.027	0.040	0.073	

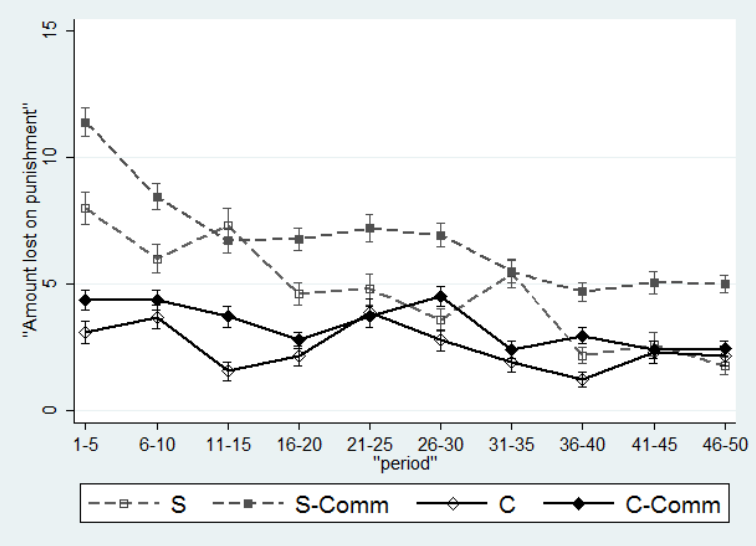
Note: OlS regression with profits regressed on treatment dummies (equation(5)). Robust standard errors clustered at the matching group level are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

the results of which are presented in Table 5. Neither the commitment coefficient nor the collectiveness coefficient are now significant. However both the sum of the of 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 a significant effect on profits (taken on their own). The combination of collective and committed punishment leads to significantly higher payoffs than standard punishment with commitment and significantly higher payoffs than collective punishment without punishment.*

It is interesting to note that while the ability to commit to punishment has a positive effect on contribution levels (as compared to standard punishment), only the combination of collective and commitment results in higher payoffs. The main reason for this is that standard 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 5 plotting the surplus loss due to punishment (allocated and received). While subjects loose approximately 5 points in the last rounds under standard punishment with commitment, the surplus loss is approximately half of that in the other treatments. An alternative way of analysing 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

Figure 5: Surplus loss due to punishment



less wasteful punishment. This fraction is 0.83 for standard punishment (0.86 across the last 10 periods) with commitment, 0.87 for standard 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% 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 (with and without commitment) thus results in less surplus lost due to punishment.

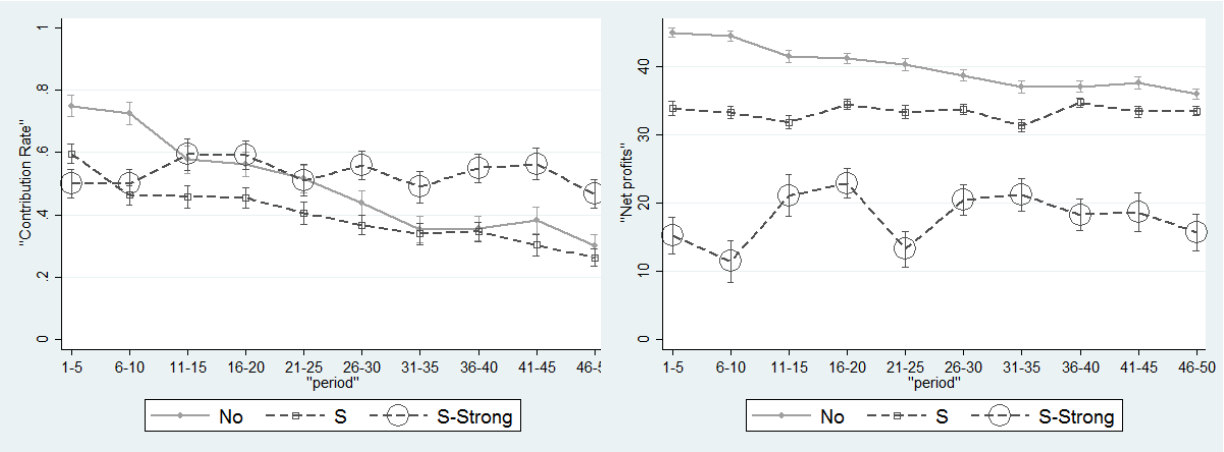
4.4 Inefficacy of Strong Punishment

It has been suggested by Ambrus and Greiner (2012) that severe peer-to-peer punishment may result in relative high contribution and welfare levels under imperfect monitoring. In their *s*trong punishment treatment subjects may subtract more punishment points from the punished subjects at a lower cost. Importantly, their framework differs from the present work in featuring a fairly low noise rate.[17] In particular, their setting featured a one sided noise rate[18] of 10%, thus making the correct assessment of one's opponent's action the rule rather than the exception.

[17]Further, their work features groups of three subjects and a more fine-tuned punishment technology where subjects can vary the severity of punishment. In contrast, in our setting subjects can either punish or not, thus enabling a straightforward implementation of commitment. In addition and in contrast to Ambrus and Greiner (2012), per round payoffs cannot not be negative in our setting.

[18]That is non-contributors were always identified as non-contributors and contributors were incorrectly labelled as non-contributors in 10% of all cases.

Figure 6: Contributions and payoffs under strong punishment



In order to assess the role of strong punishment in the present high noise environment we ran a standard peer to peer punishment treatment similar to the one in Ambrus and Greiner (2012), where subjects could subtract 30 punishment points at the cost of 5 from each other. The treatment is identical to treatment **S** except for the fact that instead of the 1:3 punishment technology now a 1:6 punishment technology is used. We ran one session with this treatment with 24 participants in six groups.

Figure 6 reports contribution rates and resulting payoffs. It can be seen that, while contributions (left panel) are slightly higher under strong punishment, payoffs (right panel) are substantially lower. Both of these differences are statistically significant. T-tests for differences in means show p-values of $p < 0.05$ for contributions across all periods and $p < 0.01$ for the last 10 periods. For payoffs the corresponding p-values are $p < 0.0001$ across all periods as well as the last 10 periods. The profit-ratio discussed in Section 4.3 equals only 0.34 in the case of strong punishment and again this is highly statistically different from that of the standard punishment treatments.

Result 7. *While contribution levels under strong punishment are higher than those under standard punishment or in the absence of the punishment, payoff levels are substantially lower than in the standard- or no- punishment case.*

One of the conclusions of Ambrus and Greiner (2012) is that the negative welfare implications associated with standard punishment under imperfect monitoring may be offset by increasing the severity of punishment, resulting in welfare levels comparable to those where no punishment is available. In the case of high noise levels this is, however, not the case. Simply providing subjects with a larger stick may result in vendetta-like dynamics, featuring excessive punishment and substantially lower welfare levels than those under standard punishment. Indeed we identify a substantial positive correlation between punishments in periods t and $t-1$ ($\rho = 0.5020$, $p = 0.0007$).

5 Conclusion

In the present paper 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 maximising equilibrium of this game. While our paper makes an important step towards understanding the efficacy and workings of collective punishment, we believe there are several important dimensions which go beyond its scope.

A key prerequisite for the effectiveness of collective sanctions is the ability to commit to punishment before the collaborative effort. Without the ability to commit to collective punishment subjects 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. 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 certain organisational structures such as workers' cooperatives or self-managing work teams but is not an accurate description of other organisations 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.[19]

References

Abbink, K., and A. Sadrieh (2009): "The pleasure of being nasty," *Economics Letters*, 105(3), 306–308.

Ambrus, A., and B. Greiner (2012): "Imperfect public monitoring with costly punishment: An experimental study," *American Economic Review*, 102(7), 3317–3332.

[19]In the absence of collective punishment Carpenter, Robbett, and Akbar (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.

Ambrus, A., and B. Greiner (2017): "Democratic punishment in public good games with perfect and imperfect observability," *mimeo*.

Bereby-Meyer, Y., and A. E. Roth (2006): "The speed of learning in noisy games: partial reinforcement and the sustainability of cooperation," *American Economic Review*, 96(4), 1029–1042.

Bornstein, G., and O. Weisel (2010): "Punishment, cooperation, and cheater detection in noisy social exchange," *Games*, 1(1), 18–33.

Brandts, J., and D. J. Cooper (2006): "A change would do you good.... An experimental study on how to overcome coordination failure in organizations," *American Economic Review*, 96(3), 669–693.

Carpenter, J., A. Robbett, and P. A. Akbar (2017): "Profit Sharing and Peer Reporting," *Management Science*.

Dal Bo, P., A. Foster, and L. Putterman (2010): "Institutions and Behavior: Experimental Evidence on the Effects of Democracy," *American Economic Review*, 100, 2205–2229.

Dickson, E. S. (2007): "On the (in) effectiveness of collective punishment: An experimental investigation," Discussion paper, New York University.

Dreber, A., D. G. Rand, D. Fudenberg, and M. A. Nowak (2008): "Winners dont punish," *Nature*, 452(7185), 348–351.

Fehr, E., and S. Gächter (2000): "Cooperation and punishment in public goods experiments," *American Economic Review*, pp. 980–994.

——— (2002): "Altruistic punishment in humans," *Nature*, 415(6868), 137–140.

Feri, F., B. Irlenbusch, and M. Sutter (2010): "Efficiency gains from team-based coordination - large scale experimental evidence," *American Economic Review*, 100(4), 1892–1912.

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

Fischer, S., K. Grechenig, and N. Meier (2016): "Monopolizing Sanctioning Power under Noise Eliminates Perverse Punishment but does not increase cooperation," *Frontiers in Behavioural Neuroscience*, 10, 1–11.

Fudenberg, D., D. G. Rand, and A. Dreber (2012): "Slow to anger and fast to forgive: cooperation in an uncertain world," *American Economic Review*, 102(2), 720–749.

Gächter, S., E. Renner, and M. Sefton (2008): "The long-run benefits of punishment," *Science*, 322(5907), 1510–1510.

Grechenig, K., A. Nicklisch, and C. Thöni (2010): "Punishment despite reasonable doubt – a public goods experiment with sanctions under uncertainty," *Journal of Empirical Legal Studies*, 7(4), 847–867.

Greiner, B. (2004): "An online recruitment system for economic experiments," .

Holmström, B. (1982): "Moral hazard in teams," *The Bell Journal of Economics*, pp. 324–340.

Kosfeld, M., A. Okada, and A. Riedl (2009): "Institution formation in public goods games," *American Economic Review*, pp. 1335–1355.

Ledford, J. G. E., I. E. E. Lawler, and S. A. Mohrman (1995): "Reward innovations in Fortune 1000 companies," *Compensation & Benefits Review*, 27(4), 76–80.

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

Mellizo, P., J. Carpenter, and P. H. Matthews (2017): "Ceding control: an experimental analysis of participatory management," *Journal of the Economic Science Association*, 3(1), 62–74.

Nalbantian, H. R., and A. Schotter (1997): "Productivity under group incentives: An experimental study," *The American Economic Review*, pp. 314–341.

Ockenfels, A., D. Sliwka, and P. Werner (2014): "Bonus Payments and Reference Point Violations," *Management Science*, 61(7), 1496–1513.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants with and without a Sword: Self-governance Is Possible.," *American Political Science Review*, 86(02), 404–417.

Patel, A., E. Cartwright, and M. Van Vugt (2010): "Punishment cannot sustain cooperation in a public good game with free-rider anonymity," Working paper in economics, University of Gothenburg.

Riedl, A., I. M. Rohde, and M. Strobel (2015): "Efficient coordination in weakest-link games," *The Review of Economic Studies*, 83(2), 737–767.

Sutter, M., S. Haigner, and M. G. Kocher (2010): "Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations," *Review of Economic Studies*, 77(4), 1540–1566.

Van Huyck, J. B., R. C. Battalio, and R. O. Beil (1990): "Tacit coordination games, strategic uncertainty, and coordination failure," *American Economic Review*, pp. 234–248.

Weber, R. A. (2006): "Managing Growth to Achieve Efficient Coordination in Large Groups," *American Economic Review*, 96(1), 114–126.

Online Appendix for "Collective Incentives and Cooperation in Teams with Imperfect Monitoring"

F.Mengel, E. Mohlin and S. Weidenholzer

Contents

A Further Theoretical Analysis of Standard Punishment with Commitment

In this appendix analyses the case of standard punishment with commitment, as implemented in the experiment. In the first stage each player j has committed $p_{ji} \in \{0, \bar{p}\}$ punishment points to player i, to be realised in case j receives a signal that i has not contributed, i.e. if j observes the signal $s_{ji} = 0$. In the second stage, each player i decides on a contribution $g_i \in \{0, \bar{g}\}$. Consider the contribution decision of individual i. Her utility is

$$u_i(g_i, g_{-i}) = \max\left\{0, \alpha G + m - g_i - P\sum_{j\neq i}^{n} \mathbb{1}[s_{ji} = 0]p_{ji} - \beta\sum_{j\neq i}^{n} \mathbb{1}[s_{ij} = 0]p_{ij}\right\}.$$

Let

$$q = \Pr(s_{ij} = \bar{g}|g_j = \bar{g}) = \Pr(s_{ij} = 0|g_j = 0) > \frac{1}{2}.$$

In our experiment $q = 0.6$, $\alpha = 0.5$, $\bar{g} = 20$, $m = 30$, $\bar{p} = 15$, and $\beta = 1/3$. Moreover, in the experiments on standard private punishment with commitment the following obtained. (1) Each subject could decide whether to punish each other with $\bar{p} = 15$ points or not. Thus, each player can maximally use 45 points (to the other players). (2) Committed punishment points can only be conditioned on the noisy signals, not on total contributions. (3) The punishment points committed towards a player is known by that player before the contribution decisions are made. Under these assumption the following result obtains.

Proposition 1. *For all values of q there is a symmetric subgame perfect equilibrium in which no one ever contributes and no one commits to punish. Restrict attention to symmetric SPE in which everyone contributes and everyone commits to punishing iff they observe a signal of non-contribution. There does not exist such an SPE if $q = 0.6$.*

It is elementary to demonstrate the existence of the symmetric subgame perfect equilibrium in which no one ever contributes and no one ever punishes. We prove the rest of the proposition by establishing that in any second stage subgame in which everyone has committed to punish if and only they obtain a signal of non-contribution it is the case that non-contribution gives a strictly higher payoff than contribution. Formally we establish:

Lemma 1. *Suppose it is the case that all of i's co-players commit to punish if an only if they obtain a signal of non-contribution. If $q = 0.6$ then $\mathbb{E}[u_i(0, g_{-i})] > \mathbb{E}[u_i(20, g_{-i})]$ for any g_{-i} such that $\sum_{j\neq i}^{n} g_j \in \{0, 20, 40, 60\}$.*

Proof. Since all signals are independent we have

$$\begin{aligned}
\mathbb{E}[u_i(20, g_{-i})] &= (1-q)^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - 25 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad + 3q(1-q)^2 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - 10 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad + 3q^2(1-q) \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 5 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad + q^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 20 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[u_i(0, g_{-i})] &= q^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - 15 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad + 3q^2(1-q) \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad + 3q(1-q)^2 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 15 - \frac{1}{3}\sum_{j\neq i}^{n} [s_{ij}=0]p_{ij}\right\}\right] \\
&\quad + (1-q)^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 30 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}_{\{s_{ij}=0\}}p_{ij}\right\}\right],
\end{aligned}$$

Since $\sum_{j\neq i}^{n} g_j \in \{0, 20, 40, 60\}$, we consider four different main cases (all of which will contain four subcases).

Case 1. **No one else contributes.** If $\sum_{j\neq i}^{n} g_j = 0$ then

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 3q^2\left(1-q\right)\mathbb{E}\left[\max\left\{0, 5-\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\ &\quad +q^3\mathbb{E}\left[20-\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}_{\{s_{ij}=0\}}p_{ij}\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] &= 3q\left(1-q\right)^2\mathbb{E}\left[15-\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}\right] \\ &\quad +\left(1-q\right)^3\mathbb{E}\left[30-\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}\right].\end{aligned}$$

Since $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}_{\{s_{ij}=0\}}p_{ij} \in \{0, 5, 10, 15\}$, there are four subcases to consider.
(1.1) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij} = 15$; we have, if $q = 0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 5q^3 = 1.08 \\ &> 0.96 = 15\left(1-q\right)^3 = \mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

(1.2) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij} = 10$; we have, if $q = 0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 10q^3 = 2.16 \\ &< 2.72 = 15q\left(1-q\right)^2 + 20\left(1-q\right)^3 = \mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

(1.3) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij} = 5$; we have, if $q = 0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 15q^3 = 3.24 \\ &< 4.48 = 30q\left(1-q\right)^2 + 25\left(1-q\right)^3 = \mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

(1.4) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij} = 0$; we have, if $q = 0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 15q^2\left(1-q\right) + 20q^3 = 6.48 \\ &> 6.24 = 45q\left(1-q\right)^2 + 30\left(1-q\right)^3 = \mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

Putting (1.1)-(1.4) together we have

$$\begin{aligned}
& \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] - \mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] \\
= \; & q^3 \mathbb{E}\left[u_i\left(20, g_{-i}\right) - u_i\left(0, g_{-i}\right) \middle| \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij} = 15\right] \\
& +3q^2\left(1-q\right)\mathbb{E}\left[u_i\left(20, g_{-i}\right) - u_i\left(0, g_{-i}\right) \middle| \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij} = 10\right] \\
& +3q\left(1-q\right)^2\mathbb{E}\left[u_i\left(20, g_{-i}\right) - u_i\left(0, g_{-i}\right) \middle| \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij} = 5\right] \\
& +\left(1-q\right)^3\mathbb{E}\left[u_i\left(20, g_{-i}\right) - u_i\left(0, g_{-i}\right) \middle| \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij} = 0\right]
\end{aligned}$$

For $q = 0.6$ we use the calculations from above to obtain

$$\begin{aligned}
& \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] - \mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] \\
= \; & \left(0.6\right)^3\left(1.08 - 0.96\right) \\
& +3\left(0.6\right)^2\left(1-0.6\right)\left(2.16 - 2.72\right) \\
& +3\left(0.6\right)\left(1-0.6\right)^2\left(3.24 - 4.48\right) \\
& +\left(1-0.6\right)^3\left(6.48 - 6.24\right) \\
= \; & -0.55776
\end{aligned}$$

Thus if $\sum_{j\neq i}^{n} g_j = 0$ and $q = 0.6$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
The remaining cases, defined by $\sum_{j\neq 1}^{n} g_j$ being equal to 20, 40, and 60 respectively, are completely analogous.
Case 2. If $\sum_{j\neq i}^{n} g_j = 20$ and $q < 0.64268$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
Case 3. If $\sum_{j\neq 1}^{n} g_j = 40$ and $q < 0.6042$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
Case 4. If $\sum_{j\neq 1}^{n} g_j = 60$ and $q < 0.60422$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
Combining cases 1-4 we find that if $q = 0.6$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$. ∎

B Sample Instructions

We provide the Instructions for treatment C-COMM. Instructions for other treatments are available on authors' webpages or upon request.

General information

You are about to participate in a decision making experiment. If you follow the instructions carefully, you can earn a considerable amount of money depending on your decisions and the decisions of the other participants. Your earnings will be paid to you in cash at the end of the experiment.

This set of instructions is for your private use only. During the experiment you are not allowed to communicate with anybody. In case of questions, please raise your hand. Then we will come to your seat and answer your questions. Any violation of this rule excludes you immediately from the experiment and all payments.

Throughout the experiment you will make decisions about amounts of tokens. At the end of the experiment all tokens you have will be converted into pounds at the exchange rate 1 pounds for 150 token and paid you in cash in addition to the show-up fee 2.50 pounds.

All your decisions will be treated confidentially both during the experiment and after the experiment. This means that none of the other participants will know which decisions you made.

Experimental Instructions

The experiment will consist of 50 decision making periods. Each period consists of two stages. At the beginning of the experiment, you will be randomly matched with three other people in this room. Therefore, there are 4 people, including yourself, participating in your group. You will be matched with the same people during the entire experiment. None of the participants knows who is in which group.

In each period you, and each other person in your group, will be given an endowment of 20 tokens. In each period you will be asked to either place your endowment in a private account or a group account.

Your private account already has 10 tokens in each period. If you place your endowment in the private account, the private account will have 30 tokens at the end of the period. If you do not place your endowment in the private account, the private account will have 10 tokens at the end of the period. This means that the private account has a return of 1. Nobody except yourself benefits from your private account. The tokens

that you place in the group account are summed together with the tokens that the other three members of your group place in the group account. Every member of the group benefits equally from the tokens in the group account. Specifically, the total amount of tokens placed in the group account by all group members is doubled and then is equally divided among the four group members. Hence, your share of the group account at the end of the first stage is

2∗ (sum of tokens in the group account) / 4

(PLEASE TURN OVER) ■

Before you decide whether to allocate your tokens to the private or to the group account, you will be asked to choose whether you would like to introduce a subtraction mechanism. Specifically, you will be asked to make four choices: whether you would like to introduce the mechanism if the total number of tokens in the group account is 0, 20, 40 or 60. All group members will make this choice simultaneously. However, only the choice of one randomly selected group member will be relevant. This means that your choice is relevant with a 25 percent chance.

If the subtraction mechanism is introduced, then it will automatically subtract tokens from all group members in case there is at least one group member who did not place their tokens in the group account. In this case it will subtract 15 tokens from each group member. In addition, if the mechanism is activated, this has a cost of 5 tokens per group member. The following table illustrates the relation between your cost in tokens and the amount of tokens that are taken away from every member of your group (including you):

Tokens subtracted	Cost for you
15	5

If all group members place their tokens in the group account, then the subtraction mechanism does not subtract tokens from anyone and there is no cost. You will know whether the subtraction mechanism was implemented for each of the four cases before you make your choice. In the second stage of each period you will be informed for each group member whether they placed their tokens in the private or group account. However, for each group member, this information is only correct with a 60% chance. It is wrong with a 40% chance.

You will also receive one piece of information that is always correct. In particular we will tell you how much money was placed in the group account.

At the end of each period, you will be informed about

- the size of the group account.

- your share of the group account (remember it is the same for all group members).
- the size of your private account.
- whether tokens were subtracted from you.
- your total earnings in this period.

This information is always correct.

All other participants will receive exactly the same instructions. Your total income in the end of the experiment is equal the sum of earnings you obtained in each period. At the end of the experiment there will be a short questionnaire for you to fill in.

This is the end of the instructions. If you have any questions please raise your hand and an experimenter will come by to answer them

C Sample Properties

Table 7 contains some balancing checks. There are about 65 percent women in our experiment, but there are no substantial nor statistically significant differences across our treatments. The average age of our participants is between 21.5-23 years across our main treatments, again without substantial nor statistically significant differences. We asked participants about which class they feel most they belong to. Around 36-38 percent of participants associate with working class (class 1), 34-35 percent with lower middle class (class 2), 24-25 percent with upper middle class (class 3) and only 0-3 percent with upper class (class 4). There are no systematic differences across treatments with the exception of treatment **C** where somewhat fewer participants identify as upper middle class. Also note that the R2 is low in all regressions. Treatments explain less than 3 percent of the variation in these parameters.

Table 7: Balancing checks

	(1) gender	(2) age	(3) class1	(4) class2	(5) class3	(6) class4	(7) risk	(8) trust	(9) rec	(10) recn
S	-0.052	0.948	-0.104	0.135	-0.020	-0.010	-0.385	-1.104*	-0.885*	-1.667
	(0.0929)	(1.138)	(0.094)	(0.088)	(0.064)	(0.035)	(0.240)	(0.624)	(0.493)	(1.377)
S-COMM	-0.014	0.014	-0.059	0.102	-0.010	-0.031	-0.510**	-0.916*	-0.488	-1.804
	(0.077)	(1.055)	(0.092)	(0.087)	(0.067)	(0.029)	(0.243)	(0.434)	(0.440)	(1.164)
C	0.055	0.252	0.125	0.002	-0.115*	-0.012	-0.127	-1.053*	-0.031	-1.500
	(0.081)	(1.334)	(0.113)	(0.097)	(0.062)	(0.034)	(0.261)	(0.544)	(0.511)	(1.260)
C-COMM	-0.091	-0.692	-0.048	0.102	-0.054	0.001	-0.107	-1.079***	-0.401	-0.196
	(0.080)	(0.983)	(0.087)	(0.070)	(0.056)	(0.034)	(0.258)	(0.399)	(0.432)	(1.265)
Constant	0.656***	22.09***	0.375***	0.344***	0.250***	0.031	6.531***	11.69***	17.78***	9.500***
	(0.061)	(0.874)	(0.077)	(0.061)	(0.044)	(0.029)	(0.180)	(0.326)	(0.360)	(1.109)
Observations	316	316	316	316	316	316	316	316	316	316
R-squared	0.011	0.009	0.023	0.010	0.009	0.009	0.013	0.015	0.015	0.023

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

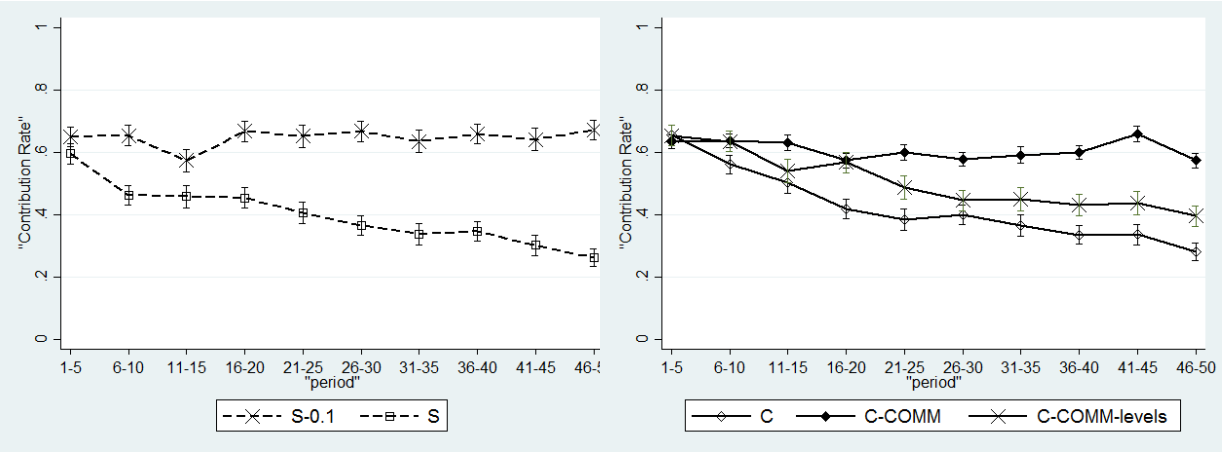
Note: OLS regression of questionnaire variables on treatment dummies. The baseline is the no punishment (**N**) treatment. Robust standard errors clustered at the matching group level are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

We also elicited attitudes towards risk, trust, positive reciprocity (rec) and negative reciprocity (recn) at the end of the experiment. Here we do see some significant differences. Most notably, participants in all our treatments with punishment possibilities (**S**, **S-COMM**, **C** and **C-COMM**) are less trusting compared to treatment **N**. Note, though, that as this outcome is elicited at the end of the experiment it is not exogenous to the treatments. It is plausible that participants who experienced punishment report to be less trusting afterwards. We should also note that there is no significant difference in trust across our punishment treatments

D Additional Treatments and Results

At the beginning of this experimental study we also conducted a standard punishment treatment with a lower noise level of 0.1 (**S-0.1**). We had 44 participants (11 groups) in this treatment. Unlike some of the earlier literature, we found, however, that in our setting cooperation rates were high under standard punishment when noise levels are low. Hence, we found that with low noise levels standard punishment was successful in solving the free-rider problem. As we are interested in situations where standard punishment fails to solve this problem, we decided subsequently to conduct the entire experiment with high noise levels. The left panel in Figure 7 compares contribution rates under standard punishment with low (0.1) and high (0.4) noise.

Figure 7: Contribution Rates Additional Treatments



Apart from this treatment we conducted a few more treatments. After we finished all sessions, we were wondering how effective C-COMM would be if we allowed participants to condition the punishment mechanism on more than just total payoffs (assume they could observe whether 1,2, or 3 people contributed). We conducted such a hybrid treatment and found that cooperation rates lie in between those of **C** and **C-COMM** (right panel of Figure 7). Last, we also conducted some sessions with no as well as standard punishment where participants were given the wrong information. We are not reporting on those but are happy to share results upon request.