Fischer, Sven; Grechenig, Kristoffel; Meier, Nicolas

Working Paper

Cooperation under punishment: Imperfect information destroys it and centralizing punishment does not help

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2013/6

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Fischer, Sven; Grechenig, Kristoffel; Meier, Nicolas (2013) : Cooperation under punishment: Imperfect information destroys it and centralizing punishment does not help, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2013/6, Max Planck Institute for Research on Collective Goods, Bonn

This Version is available at:
https://hdl.handle.net/10419/84987

**Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2013/6**

Cooperation under punishment: Imperfect information destroys it and centralizing punishment does not help

Sven Fischer
Kristoffel Grechenig
Nicolas Meier

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2013/6

Cooperation under punishment: Imperfect information destroys it and centralizing punishment does not help

Sven Fischer / Kristoffel Grechenig / Nicolas Meier

April 2013

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Cooperation under punishment:

Imperfect information destroys it and centralizing punishment does not help

Sven Fischer Kristoffel Grechenig* Nicolas Meier†

27.02.2013

Max Planck Institute for Research on Collective Goods
University of Bonn

We run several experiments which allow us to compare *cooperation* under *perfect* and *imperfect information* and under a *centralized* and *decentralized punishment* regime. We find that (1) centralization by itself does not improve cooperation and welfare compared to an informal, peer-to-peer punishment regime and (2) centralized punishment is equally sensitive to noise as decentralized punishment, that is, it leads to significantly lower cooperation and welfare (total profits). Our results shed critical light on the widespread conjecture that the centralization of punishment institutions is welfare increasing in itself.

Keywords: cooperation, public goods, centralized punishment, imperfect information, decentralized punishment, peer to peer punishment
JEL: K42, H42, C92, D03

*Sven Fischer and Kristoffel Grechenig (corresponding authors): Max Planck Institute for Research on Collective Goods, Bonn, Germany; fischer@coll.mpg.de; +49/228/91416-53; grechenig@coll.mpg.de; +49/228/91416-51
†Current affiliation: Universität Köln and Max Planck Institute for Research on Collective Goods, Bonn, Germany, meier@coll.mpg.de

1. Introduction

Modern societies have centralized the sanctioning power as a means to enforce norms (Weber 1919). This monopoly has often been justified on the premise that private, decentralized enforcement has (higher) negative externalities (Clotfelter 1978, Polinsky 1980). In the extreme, this argument interprets public enforcement as part of a social contract that is necessary to prevent anarchy that comes with a war of everyone against everyone (bellum omnium contra omnes, Hobbes, 1642, 1651).

Experimental research allows to trace the centralization of punishment back to its roots by establishing a simple environment of social interaction. In such an environment, we can isolate the centralization of punishment *per se*, abstracting from the different institutional features of centralized punishment that subsequently followed throughout history.

Experiments repeatedly demonstrated that decentralized, informal, peer-to-peer punishment increases cooperation (Yamagishi 1986, Ostrom et al. 1992, Fehr and Gächter 2000, Fehr and Gächter 2000) and welfare in the long run (Gächter et al. 2008), compared to an environment *without punishment*. Various studies challenge the robustness of these results (for a recent overview, see Nikiforakis 2013): for example, on the basis of punishment that is targeted at cooperators, referred to as anti-social punishment (Herrmann et al. 2008), on the basis of counter-punishment (Nikiforakis 2008, Nikiforakis et al. 2012) and on the basis of a non-trivial degree of noise regarding contributions, where cooperation decreases significantly and total earnings drops below what is achieved under a regime without any punishment (Grechenig et al. 2010, Ambrus and Greiner 2012).

Recent experimental studies test the effectiveness of *formal, centralized* enforcement mechanisms compared to informal, decentralized regimes, thereby capturing important aspects of institutions. They suggest that centralized punishment has a positive effect on cooperation and welfare, and hence support the conjecture that centralized punishment has emerged to overcome social dilemmas. The results are based on the idea that centralized, formal punishment makes use of the positive incentive effects of punishment, while preventing counter-punishment, anti-social punishment, and other negative effects that come from a system of private, decentralized, peer-to-peer punishment. One strand of literature characterizes centralization as a mechanism allowing to commit to a sanctioning scheme, such that punishment is automatically carried out when certain conditions are met. Such a sanctioning scheme is either exogenous (Andreoni and Gee 2012, Kube and Traxler 2011) and/or determined according to some exogenous voting rule (Kosfeld

and Riedl 2004, Tyran and Feld 2006, Guillen et al. 2007, Sutter et al. 2010, Putterman et al. 2011, and also Andreoni and Gee 2012). Applied to institutions, both approaches implicitly view centralization as a commitment mechanism and thus assume that commitment through institutions is possible (cf. Bowles 2003). Other studies allow for an actual person to carry out the punishment without prior commitment. The enforcer is either randomly chosen (Fehr and Fischbacher 2004, Nelissen and Zeelenberg 2009, O'Gorman et al. 2009, Engel and Irlenbusch 2010, Leibbrandt and López-Pérez 2011, Leibbrandt and López-Pérez 2012), or elected according to some exogenous voting rule (Baldassarri and Grossman 2011). Whether positive effects result from commitment, self-control through elections, and other considerations, or from the centralization *per se* remains widely unanswered. Arguably, if commitment through informal punishment was perfect, it could be just as effective as centralized punishment with commitment.

In order to test the effect of centralization *per se*, we hold all other considerations constant across institutions (treatments): particularly, (1) punishers cannot commit to punishment ex ante, (2) contributors cannot deliberately withdraw punishment power from some or assign it to others (e.g. through voting), (3) the direct consequences from punishment are the same, and (4) there are no differences in externalities resulting from punishment.

By abstracting from institutional factors, we return to the origins of formal punishment as a centralization of informal sanctioning regimes (Turnbull 1962, Sahlins 1972, Guala 2012). Since previous studies suggest that the negative effects of decentralized punishment are particularly pronounced under imperfect information (Grechenig et al. 2010, Ambrus and Greiner 2012), we test whether centralized punishment is less sensitive to noise. To the best of our knowledge, we are the first to analyze the effects of centralization of punishment *per se* on cooperation and to study centralized punishment under *imperfect information*.

We find that centralized punishment is *highly sensitive to imperfect information* as is decentralized punishment, both with respect to cooperation rates and total earnings: cooperation and earnings are considerably lower under imperfect information. Regarding incentives created by decentralized and centralized punishment institutions, we find that central punishers care about the *absolute* level of cooperation, while peer-to-peer punishers (decentralized) only care about the *relative* cooperation behavior. We also find that under centralized punishment, cooperative participants tend to decrease their contributions. This may result from the fact that the participants cannot react to group differences in contributions by applying punishment. Under decentralized punishment,

we observe more *anti-social* punishment, meaning that low contributors punish cooperative types, who then decrease their contributions in the following period.

Our results put into perspective findings from recent studies that emphasize the importance of centralized punishment, as well as the conjecture that centralized punishment may be less sensitive to noise.

The remainder of the paper is organized as follows. In the following chapter, we describe the experimental game and design in detail, before we present and analyze experimental behavior in section 3. Section 4 concludes with a discussion of our findings.

2. Experimental Design

We use a standard finitely repeated linear public-goods game with a voluntary contribution mechanism. Participants interact in groups of five over 30 periods in a partner design, where every period has two stages, a contribution and a punishment stage. In our set of experiments we have two treatments with two conditions each (2×2 design). In the first dimension, we compare two different punishment *institutions*: Decentralized (DEC) and Centralized punishment (CEN), in the second, we contrast a perfect with a noisy signal of contributions, indicated by parameter λ with $\lambda = 1$ or $\lambda = 0.5$.

Four of the five participants ($i \in \{1..4\}$) in each group can contribute to the public good; the remaining participant, the so-called *Authority* (A), benefits from the public good but cannot contribute himself. In treatments with centralized punishment, the authority decides over punishment; in decentralized punishment treatments, the additional participant is merely passive, that is, the participant cannot make any decision but is, nevertheless, affected by the contribution and punishment decisions of the four others.

After the contribution decision, all five receive perfect or imperfect signals about the contribution decisions of the four participants $i = 1, .., 4$, according to the condition of the *information* treatment. Then they can apply punishment according to the condition of the *institution* treatment (CEN vs. DEC).

2.1. Stage I

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

In line with the overwhelming majority of public goods experiments, we chose a marginal per capita return of 0.4. Hence, the monetary payoff of player i in the first

stage is given by

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

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

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

2.2. Stage II

In the second stage each subject, including the authority A, receives a signal s_k about the contribution of subject k with

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

The signal $\widetilde{g}_k$ is a group wise realization out of the uniform distribution $\{0, 2, 4, ..., 20\} \setminus \{g_k\}$. All participants receive the same signals about contributions of others, which ensures that the information of punishers is kept constant across treatments. However, they do not know which signal others receive about their own contribution. To reduce possible identification of group members (via contributions over time), every period each of the four subjects was randomly given a number between 1 and 4. We contrast a perfect signal ($\lambda = 1$) with a noisy signal where $\lambda = 0.5$. Under $\lambda = 0.5$, a contribution of 6 would lead to the signal "6" with probability .5, and to any other signal with probability .05 (.5/10).

2.2.1. Punishment in DEC

In treatment DEC, each of the four regular participants i receives a punishment endowment of $e_p = 10$ punishment points, and authority A receives an additional endowment of $e_p^A = 40$. The four regular participants can distribute punishment points, where each point costs them one unit and also reduces the authority's income by one. At the same time every received punishment point reduces the target participant's income by three units. More specifically, denoting a punishment point sent by i to j with p_{ij}, the total payoff of subject i is:

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

and the payoff of participant A is:

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

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

2.2.2. Punishment in CEN

Participants receive the same endowment of $e_p = 10$ and $e_p^A = 40$, respectively. However, in CEN, only authority A can distribute punishment points. Every punishment point distributed by A reduces A's payoff by one, the punished subject's payoff by three, and the payoff of each other participant by 1/3. This keeps the overall costs of punishment constant across treatments (participants finance the punishment applied by A, except one's own punishment).[1] Thus, in CEN, final payoffs are determined as follows:

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

and

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

where equivalently to p_{ij} we denote with p_{Aj} the number of punishment points assigned to j by A. This payoff structure keeps all considerations outside a *per se* centralization constant.

2.3. Setup

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

The experiments were run in the experimental laboratory of the University of Bonn (EconLab) in March 2012 and was programmed and conducted with the software z-Tree (Fischbacher 2007). We ran a total of 8 sessions with 160 participants, mostly

[1] Alternatively, one could exclusively burden the punisher with the costs. However, with respect to external validity our design is considerably more realistic, as in modern societies the costs of punishment institutions are shared by everyone.

undergraduate students, divided into 32 groups, and no participant took part in more than one session. We relied on ORSEE (Greiner 2004) for recruiting.

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

3. Results

3.1. Institutional Incentives

We look at the institutional incentives under perfect/imperfect information and we test whether there are differences between institutions. Under perfect information, punishment can be conditioned on the contribution behavior; under imperfect information, contribution behavior must be inferred from noisy signals. We estimate the linear mixed effect model, $y_{git} = x'_{git}\beta + u_g + u_i + e_{git}$, where g indexes the matching group and i the subject nested in g.[2] When estimating the effects of contribution decisions on subsequent punishment, we use individual deviations from group averages, as group averages are a widely accepted norm both in the lab and in the field.

Figure 1 and Table 1, models (1) and (2) show that lower contributions (relative to average contributions) lead to stronger punishment. This effect is substantial under perfect information and sets sufficient incentives to cooperate, as with 0.963 the coefficient on *dev* is way above 0.6 (a reduction of 1 unit in contributions yields an increase in profits of 0.6, but it is punished by 0.963). Under imperfect information, the effect is close to zero and clearly too weak to create sufficient incentives ($dev = 0.092 < 0.6$). Coefficients avC show that lower levels of average contributions lead to stronger punishment in centralized institutions, while this effect is not present in DEC ($-0.545 + 0.573$ is not different from zero, Wald test, $p = 0.5606$). This supports the conjecture that the authority is more "efficiency-minded", while decentralized punishers care more about relative payoffs.

Models (3) and (4) report results from positive and negative deviations from group averages separately, i.e., we estimate two separate regression lines, one for positive and one for negative deviations. The regression lines in Figure 1 illustrate the results. Model (3) shows the regression results for perfect signals. As before we find a significant punishment of participants who contribute less than average, which significantly increases

[2] We assume that the group - (u_g) and subject (u_i) effects are independent of the fixed effects (random effects estimation) and estimate via restricted maximum likelihood.

in the deviation.[3] Effects from positive deviations ($D_{dev<0}$) and their interaction effects in Model (3) show that under decentralized punishment cooperative participants who contribute more than others are increasingly punished the more they deviate. Such an increase in anti-social punishment is not present in the centralized environment (however, there is also anti-social punishment in *CEN*). This is an indication that competitive preferences play out (more strongly) through punishment behavior in the decentralized punishment institution. While punishment is costly it reduces the payoff of others even more. However, it is unclear why this behavior is targeted at participants who contribute more and therefore already earn less. It is noteworthy that in 93% of the cases where someone who contributed above average was punished, punishment was exercised by someone who contributed less than the punished participant.

Model (4) suggests that imperfect information makes it difficult to condition punishment on contributions. Negative deviations from group averages still lead to significantly more punishment. However, incentives are too weak ($0.123 < 0.6$).

Overall, our results confirm some of the conjectures about possible advantages of centralizing punishment. These effects seem less pronounced under imperfect information and, more importantly, they do not carry over to cooperation and total profits, as we show in the following section.

3.2. Cooperation & Total Profits

We test whether the centralized institution leads to higher cooperation rates and total profits, both for perfect and for imperfect information, and we compare perfect information environments with imperfect information environments for both centralized and decentralized institutions separately. Figures 2 and 3 show contributions and profits over time across treatments.

Wilcoxon rank sum tests, two-sided, show that there are no significant differences in *Contributions* when comparing decentralized with centralized punishment averaged over all periods, neither for perfect nor imperfect information (DEC/1 v. CEN/1, p = .834, DEC/.5 v. CEN/.5 , p = .753). We also find no significant difference in any of the 30 periods if we test every period separately (all p-values $> .1$).

Differences between treatments with perfect information and imperfect information are highly significant (Wilcoxon rank sum, two-sided, DEC/1 v. DEC/.5, p = .012,

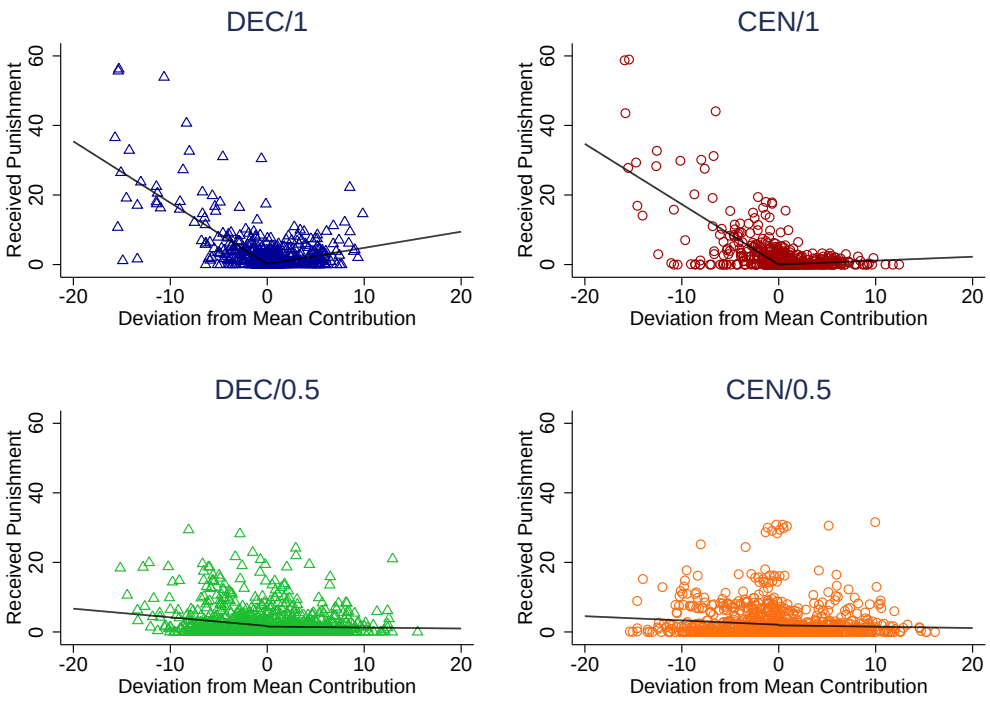
[3] See the constant and the effect on $D_{\text{dev}\leq 0}\times$dev. Also note that DEC$\times D_{\text{dev}\leq 0}\times$dev is insignificant and joint tests also confirm that the overall effect in treatment DEC is still significant.

Table 1: Received Punishment

RecPun	(1) $\lambda = 1$	(2) $\lambda = 0.5$	(3) $\lambda = 1$	(4) $\lambda = 0.5$
DEC	-9.247***	-0.702	-7.805***	-0.919
	(-5.82)	(-0.75)	(-5.18)	(-0.97)
dev	-0.963***	-0.0926***		
	(-21.40)	(-3.46)		
DEC×dev	0.0562	-0.0576		
	(0.87)	(-1.45)		
avC	-0.545***	-0.0576	-0.289***	-0.0641
	(-11.82)	(-1.48)	(-7.05)	(-1.54)
DEC×avC	0.573***	0.0778	0.479***	0.0729
	(8.62)	(1.26)	(8.36)	(1.15)
$D_{dev>0}$			0.0436	-0.171
			(0.15)	(-0.49)
$D_{dev>0}$×dev			0.111	-0.0388
			(1.38)	(-0.65)
DEC×$D_{dev>0}$×dev			0.353***	0.0109
			(3.64)	(0.14)
$D_{dev\leq 0}$×dev			-1.736***	-0.123**
			(-30.97)	(-2.17)
DEC×$D_{dev\leq 0}$×dev			-0.0287	-0.126
			(-0.38)	(-1.57)
_cons	10.48***	2.617***	4.857***	2.564***
	(9.33)	(4.08)	(4.46)	(3.96)
N*Subj.*	1920(64)	1920(64)	1920(64)	1920(64)
chi2	982.9	40.30	2423.9	46.47
p	< 0.001	< 0.001	< 0.001	< 0.001
AIC	10895	11031	10138	11038

Note: Coefficients (t-statistics) of linear mixed effect regressions including group-wise random effects and subject-wise random effects nested in group effects. Random effects and fixed effects specification are identical according to Hausman tests. *DEC* is a dummy variable = 1 for treatments with decentralized punishment, *dev* captures the difference between one's own contribution and the average contribution in a group, *avC* captures average contributions in a group, $D_{dev>0}$ is a dummy for positive deviations from group averages (in contributions), $D_{dev\leq 0}$ is a dummy for negative deviations from group averages (in contributions). The regression lines from models (3) and (4) are illustrated in Figure 1. Interaction effects are indicated by ×, p-values are reported as * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1: Punishment Reaction to Contributions



Note: Scatter plots use jitter to illustrate overlapping points. Lines are the regression lines from estimation (3) for treatments CEN/1 and DEC/1 and (4) for treatments CEN/0.5 and DEC/0.5 in Table 1, for treatment averages of variable avC (average contribution).

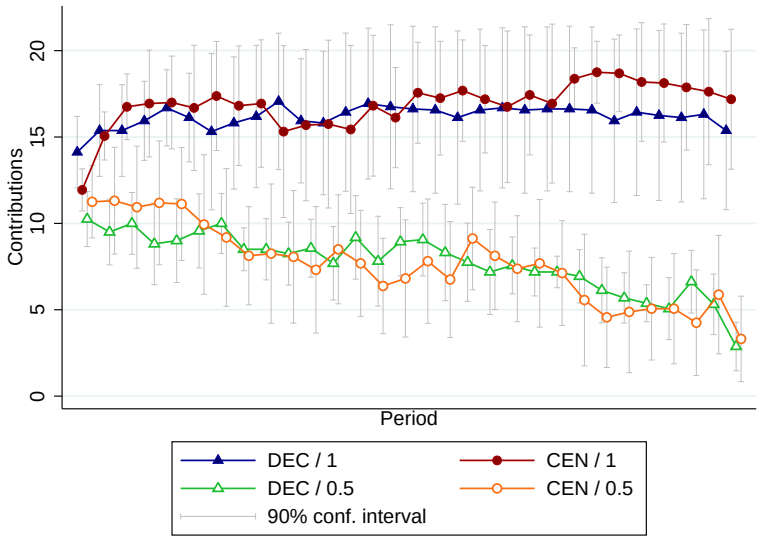
CEN/1 v. CEN/.5, p = .009).[4] Differences are also significant for every single period if tested separately, with p-values $< .05$, except for the first two periods of CEN/1 vs. CEN/.5.

Results for *Profits* are similar to those for contributions. Differences of average profits over all periods between centralized and decentralized punishment institutions are insignificant at both noise levels (DEC/1 v. CEN/1, p = .674, DEC/.5 v. CEN/.5 , p = .916, Wilcoxon rank sum test), and for every single period, if tested separately with p-values $> .1$ (except in period 28 for DEC/.5 v. CEN/.5).

Differences between treatments with perfect information and imperfect information are partly only weakly significant if tested for all periods (Wilcoxon rank sum, two-sided, DEC/1 v. DEC/.5, p = .059, CEN/1 v. CEN/.5, p = .003). Separate tests for every single period show that treatment differences emerge over time, such that differences are

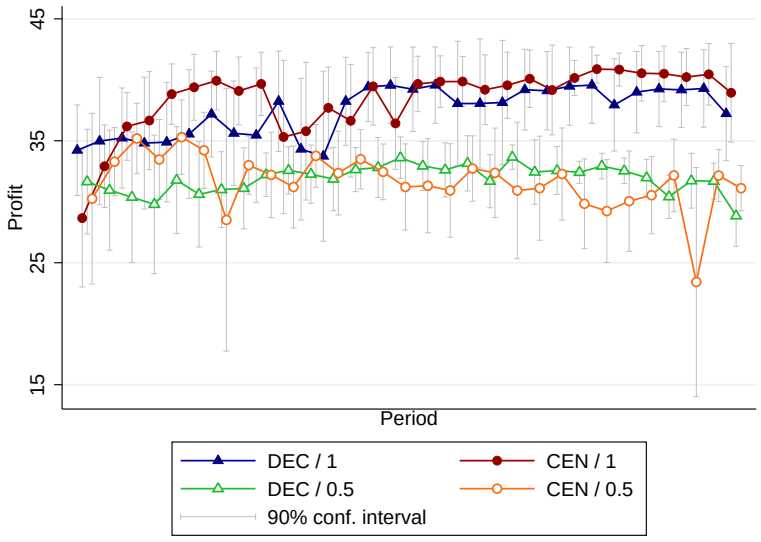
[4]Please note that the confidence intervals in Figures 2 and 3 are based on parametric comparisons. Slight differences to the nonparametric results reported here are therefore unsurprising.

Figure 2: Contributions



Note: The confidence intervals are based on the distribution of averages per matching group.

Figure 3: Total Profits



Note: The confidence intervals are based on the distribution of averages per matching group.

significant for the second half of the experiment in 13 of 15 periods in DEC/1 v. DEC/.5 (p-values $< .05$), and in all 15 periods in CEN/1 v. CEN/.5 (p-values $< .05$).

We obtain similar results if we test for treatment differences using parametric tests.

More specifically, random effects regressions with treatment dummies allow the same inferences we made using non-parametric tests.

Our results suggest that centralized institutions are just as sensitive to imperfect information as decentralized institutions are. While the literature has pointed to this effect under decentralized punishment, the 'hope' was that centralized punishment may be less sensitive. Out findings support the conjecture that the alleged effectiveness of centralized punishment is due to other factors, such as the possibility of centralized punishment to commit *ex ante* to certain punishment, voting mechanisms, etc. Centralization *per se* does not have the alleged increased effectiveness and it is highly sensitive to noise.

3.3. Reaction to punishment

In order to understand why the differences between centralized and decentralized punishment in terms of punishment behavior do not lead to different total outcomes, we further explore the reaction to punishment (see, e.g., Grechenig et al. 2010). Particularly, we analyze how contributions at period $t+1$ change, following a punishment at t. The estimations in Table 2 are mixed effects regressions of changes in individual contributions on previously experienced punishment (recPun) and a dummy indicating whether the previous contribution was above average (DhighC). Despite some inherent (potential) endogeneity issues, one can identify some important results.[5] Under perfect information (model 1), participants significantly increase their contribution the more they were punished (recPun$_{t-1}$). However, if punishment is decentralized, this effect is significantly weaker (DEC×recPun$_{t-1}$). *Anti-social punishment*, on the other hand, has no effect in CEN (where it also hardly occurs), but results in significantly less cooperation in the decentralized institution.[6] This difference in institutions suggests lower total payoffs in DEC. However, there is another important difference that favors decentralized punishment. In CEN, a high contributor tends to decrease his contribution in the following period (DhighC$_{t-1}$), an effect which is absent in DEC.[7] This may be due to the fact that high contributors may satisfy competitive preferences only by decreasing their contributions in CEN (if the authority does not apply punishment according to their sentiments), while in DEC, they may choose to punish others according to their punishment sentiments. This may explain why, despite different punishment behavior (see section 3.1), both institutions result in equal levels of total welfare.

[5] Note that both previously received punishment and the dummy DhighC$_{t-1}$ are not strictly exogenous. We were unable to find valid instruments, even in an Arelano-Bond dynamic panel data estimation.
[6] See coefficient DhighC$_{t-1}$×recPun$_{t-1}$ for CEN and DEC×DhighC$_{t-1}$×recPun$_{t-1}$ for DEC.
[7] More specifically, the effect on DEC×DhighC$_{t-1}$ offsets the one on DhighC$_{t-1}$.

Unsurprisingly, under imperfect information, there are less dynamic effects. Again high contributors reduce their contributions in CEN. Also, this effect is significantly different in DEC. However, in DEC, high contributors still significantly reduce their contributions by 4.8337 (Wald-test, $p = 0.0157$).

Table 2: Change in Contribution

ΔContribution	(1) $\lambda = 1$	(2) $\lambda = 0.5$
_cons	0.0304	2.3270***
	(0.1374)	(0.3715)
DEC	-0.1687	-0.4410
	(0.1915)	(0.5355)
recPun_{t-1}	1.2200***	-0.0968
	(0.0683)	(0.1261)
DEC$\times\text{recPun}_{t-1}$	-0.1746*	0.2884
	(0.0970)	(0.2027)
$D\text{highC}_{t-1}$	-1.4712***	-6.2251***
	(0.2652)	(0.4244)
DEC$\times D\text{highC}_{t-1}$	1.0195**	1.3914**
	(0.4139)	(0.6080)
$D\text{highC}_{t-1}\times\text{recPun}_{t-1}$	0.4822	-0.0951
	(0.6466)	(0.2494)
DEC$\times D\text{highC}_{t-1}\times\text{recPun}_{t-1}$	-2.6192***	-0.3061
	(0.6894)	(0.3645)
N$Subj.$	1856(64)	1856(64)
chi2	681.04	429.18
p	< 0.001	< 0.001

Note: Coefficients (t-statistics) of linear mixed effect regressions including group-wise random effects and subject-wise random effects nested in group effects. Dependent variable ΔContribution is the change in contribution from period $t-1$ to t. DEC is a dummy variable equal 1 for treatments with decentralized punishment, recPun_{t-1} is the number of punishment points received at $t-1$, and dummy $D\text{highC}_{t-1}$ equals 1 if the contribution in the previous period was above average. Interaction effects are indicated by $\times$, p-values are reported as * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

4. Discussion

Centralized punishment institutions have been praised in the economic, legal, and political science literature, with recent support from experimental research. Experiments with

formal, centralized punishment regimes suggest higher overall efficiency than decentralized, peer-to-peer punishment. However, it is unclear whether centralization *per se* is beneficial or whether other institutional differences drive the results. This issue carries over to studies with endogenous institutions, where centralized punishment prevails if it comes with additional advantages (Traulsen et al. 2012), but loses against decentralized punishment in a *ceteris paribus* comparison under perfect information (Grechenig et al. 2013).

Our experiment is designed to test whether centralization *per se* under *perfect* and *imperfect* information affects behavior and outcomes, when imposed exogenously, holding everything else constant. We find no differences in contributions, average punishment, and welfare. Even the patterns over time are fairly similar. This holds despite some significant differences in how contribution behavior is being punished. While centralized authorities do not punish participants who contribute more than average, we observe substantial anti-social punishment if every group member can engage in punishment in DEC. Such anti-social punishment could be driven by competitive preferences or considerations connected to delayed counter-punishment. Dynamically, this results in less effort of the punished subject in the following period, and therefore creates inefficiencies, in addition to the ones resulting from the costs of punishment. However, under centralized punishment an analysis of dynamic behavior also reveals a source of inefficiency. Participants who realize that they contributed more than average tend to reduce their contribution. This could be due to the fact that high contributors cannot apply punishment themselves. In the decentralized institution the same type of participant may choose to increase punishment instead of decreasing his contribution. This may even deter very high contributors from changing their cooperation behavior, thus, in this respect favoring decentralized over centralized punishment. These findings are consistent with studies on different kinds of decentralized punishment which could be interpreted as a partial centralization and which, despite some differences, lead to the same overall profits (Nikiforakis et al. 2010, Leibbrandt et al. 2012).

When comparing perfect information to imperfect information, we show that cooperation and total earnings are significantly lower in noisy environment. While this has been analyzed with regard to decentralized punishment (Grechenig et al. 2010, Ambrus and Greiner 2012), the hope was that centralized punishment may be less sensitive. Comparing the two punishment institutions, we find that they are highly and equally sensitive to noise, such that cooperation and total profits significantly and substantially decrease under imperfect information.

While we find substantial differences in behavior, our results clearly reject the notion that the monopolization of the punishment power *itself* has positive effects. However, both institutions have different causes and sources of inefficiencies. This raises the question whether in an interaction with other environmental aspects, one institution may become more effective in enforcing cooperation than the other.

Acknowledgements

We thank the Max Planck Society for financial support. We also thank Christoph Engel, Bruno Frey, Ben Greiner, Werner Güth, Oliver Kirchkamp, Marco Kleine, Andreas Leibbrandt, Andreas Nicklisch, Axel Ockenfels, Ro'i Zultan, the participants of the ESA meeting New York 2012 and the participants of seminars in Bonn for helpful comments and discussion.

References

Ambrus, A. and B. Greiner (2012, September). Imperfect public monitoring with costly punishment: An experimental study. *American Economic Review 102*(7), 3317–32.

Andreoni, J. and L. Gee (2012). Gun for hire: Does delegated enforcement crowd out peer punishment in giving to public goods? *Journal of Public Economics 96*(11-12), 1036–1046.

Baldassarri, D. and G. Grossman (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences of the United States of America 108*(27), 1–5.

Bowles, S. (2003). *Microeconomics: Behavior, institutions, and evolution*. Princeton: Princeton University Press.

Clotfelter, C. T. (1978). Private security and the public safety. *Journal of Urban Economics 5*(3), 388–402.

Engel, C. and B. Irlenbusch (2010). Turning the Lab into Jeremy Bentham's Panopticon - The Effect of Punishment on Offenders and Non-Offenders. *MPI Collective Goods Preprint 2010/06*.

Engel, C. and B. Rockenbach (2009). We are not alone: the impact of externalities on public good provision. Preprints of the Max Planck Institute for Research on Collective Goods 2009,29, Bonn.

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

Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review 90*(4), 980–994.

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

Gächter, S., E. Renner, and M. Sefton (2008). The long-run benefits of punishment. *Science 322*(December), 2008.

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.

Grechenig, K., A. Nicklisch, and C. Thöni (2013). Information-sensitive leviathans: The emergence of centralized punishment. *mimeo*.

Greiner, B. (2004). An online recruitment system for economic experiments. In K. Kremer and V. Macho (Eds.), *Forschung und wissenschaftliches Rechnen* (GWDG Beric ed.)., pp. 1–15. Göttingen.

Guala, F. (2012). Reciprocity: Weak or strong? what punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences 35*(01), 45–59.

Guillen, P., C. Schwieren, and G. Staffiero (2007). Why feed the leviathan? *Public Choice 130*, 115–128. 10.1007/s11127-006-9075-3.

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

Kosfeld, M. and A. Riedl (2004). The design of (de)centralized punishment institutions for sustaining cooperation. *Tinbergen Institute Discussion TI 2004-025/1*.

Kube, S. and C. Traxler (2011). The Interaction of Legal and Social Norm Enforcement. *Journal of Public Economic Theory 13*(2006), 639–660.

Leibbrandt, A. and R. López-Pérez (2011). The dark side of altruistic third-party punishment. *Journal of Conflict Resolution 55*, 761–784.

Leibbrandt, A. and R. López-Pérez (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization 84*(3), 753–766.

Leibbrandt, A., A. Ramalingam, L. Sääksvuori, and J. M. Walker (2012). Broken punishment networks in public goods games: Experimental evidence. Jena economic research papers 2012,004, Jena.

Nelissen, R. M. A. and M. Zeelenberg (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the funcitons of altruistic sanctions. *Judgement and Decision Making 4*(7), 543–553.

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

Nikiforakis, N. (2013). Self-Governance through Altruistic Punishment? In *Social Dilemmas: New Perspectives on Reward and Punishment*. New York: Oxford University Press (forthcoming).

Nikiforakis, N., H.-T. Normann, and B. Wallace (2010, January). Asymmetric enforcement of cooperation in a social dilemma. *Southern Economic Journal 76*(3), 638–659.

Nikiforakis, N., C. Noussair, and T. Wilkening (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics 96*(9-10), 797–807.

O'Gorman, R., J. Henrich, and M. Van Vugt (2009, January). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings. Biological sciences / The Royal Society 276*(1655), 323–9.

Ostrom, E., J. Walker, and R. Gardner (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review 86*(2), 404–417.

Polinsky, A. M. (1980). Private versus Public Enforcement of Fines. *The Journal of Legal Studies 9*(1), 105–127.

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

Sahlins, M. (1972). *Stone Age Economics*. Routledge Classic Ethnographies Series. New York.

Sutter, M., S. Haigner, and M. G. Kocher (2010, April). Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies 77*, 1540–1566.

Traulsen, A., T. Röhl, and M. Milinski (2012, September). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society B 279*(1743), 3716–21.

Turnbull, C. (1962). *The forest people*. Anchor books. Simon and Schuster.

Tyran, J.-R. and L. P. Feld (2006, March). Achieving Compliance when Legal Sanctions are Non-deterrent. *Scandinavian Journal of Economics 108*(1), 135–156.

Weber, M. (1919). Politics as a vocation. In H. H. Gerth and C. W. Mills (Eds.), *Essays in Sociology*. New York: Oxford University Press. Translated reprint (1958).

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology 51*(1), 110–116.

APPENDIX

A. Tests per round

(For review purposes only.)

Table 3: Contributions - Two-sample Wilcoxon rank-sum (Mann-Whitney) test, group averages

	Ho: Contribution(Treat1) = Contribution(Treat2) for each treatment 8 independent observations **p-value =**			
Period	*DEC1* vs. *CEN1*	*DEC.*5 vs. *CEN.5*	*DEC1* vs. *DEC.5*	*CEN1* vs. *CEN.5*
all	0.8335	0.7527	**0.0117**	**0.0086**
1	0.1264	0.4923	**0.0306**	0.6353
2	0.4929	0.3177	**0.0134**	0.0580
3	0.5621	0.3174	**0.0116**	**0.0153**
4	0.5609	0.3166	**0.0044**	**0.0204**
5	0.8693	0.3439	**0.0027**	**0.0300**
6	0.7422	0.6355	**0.0084**	**0.0262**
7	0.4002	0.6726	**0.0153**	**0.0189**
8	0.3238	0.7918	**0.0117**	**0.0103**
9	0.6852	0.8746	**0.0110**	**0.0105**
10	0.3983	0.6733	**0.0111**	**0.0298**
11	0.7010	0.6737	**0.0114**	**0.0170**
12	1.0000	0.6719	**0.0142**	**0.0342**
13	0.4655	0.5268	**0.0139**	**0.0266**
14	0.7017	0.4613	**0.0090**	**0.0059**
15	0.2972	0.4005	**0.0112**	**0.0061**
16	0.7014	0.3703	**0.0129**	**0.0055**
17	0.7017	0.6355	**0.0176**	**0.0055**
18	0.7826	0.3710	**0.0268**	**0.0059**
19	0.9519	0.7128	**0.0121**	**0.0047**
20	0.7713	0.7524	**0.0105**	**0.0069**
21	0.7984	0.6355	**0.0096**	**0.0076**
22	0.6095	0.3987	**0.0102**	**0.0070**
23	0.5463	0.2056	**0.0095**	**0.0012**
24	0.8358	0.2247	**0.0095**	**0.0008**
25	0.7627	0.5271	**0.0105**	**0.0014**
26	0.3008	0.8328	**0.0102**	**0.0021**
27	0.5651	0.7120	**0.0104**	**0.0025**
28	0.5651	0.1860	**0.0121**	**0.0024**
29	0.8558	0.8744	**0.0105**	**0.0046**
30	0.3978	0.8324	**0.0109**	**0.0025**

Table 4: Profits - Two-sample Wilcoxon rank-sum (Mann-Whitney) test, group averages

	Ho: Profit(Treat1) = Profit(Treat2) for each treatment 8 independent observations **p-value =**			
Period	*DEC1* vs. *CEN1*	*DEC.*5 vs. *CEN.5*	*DEC1* vs. *DEC.5*	*CEN1* vs. *CEN.5*
all	0.6744	0.9164	**0.0587**	**0.0033**
1	0.1415	1.000	0.2480	0.7527
2	0.2480	0.8336	0.2076	0.9164
3	0.7520	0.2069	0.2473	0.5984
4	0.7120	0.4942	0.0738	0.2072
5	0.2278	0.2936	0.3166	0.1247
6	0.4428	0.1556	0.0919	**0.0342**
7	0.2866	0.3717	**0.0205**	**0.0105**
8	0.2278	0.5286	0.0740	**0.0190**
9	0.1970	0.9581	0.2059	**0.0056**
10	0.5628	0.5632	0.0550	0.2042
11	0.5836	0.5990	0.5982	0.2441
12	0.5749	0.8335	0.2897	0.0721
13	0.4002	0.2926	0.0484	0.2463
14	0.5463	0.7527	**0.0172**	**0.0152**
15	0.1645	0.3181	**0.0235**	0.0917
16	0.6293	0.4619	**0.0309**	**0.0082**
17	0.8479	0.6742	**0.0081**	**0.0040**
18	0.8984	0.9580	0.0666	**0.0055**
19	0.9519	0.5992	**0.0251**	**0.0164**
20	0.8170	0.9581	**0.0142**	**0.0081**
21	0.7629	0.5990	**0.0427**	**0.0040**
22	0.9519	0.5990	**0.0327**	**0.0105**
23	0.6293	0.1031	**0.0131**	**0.0025**
24	0.8984	0.1270	**0.0131**	**0.0009**
25	0.5871	0.3717	0.1658	**0.0014**
26	0.4436	0.6731	**0.0250**	**0.0018**
27	0.7629	0.2929	**0.0056**	**0.0028**
28	0.6095	**0.0312**	**0.0077**	**0.0008**
29	0.7629	1.0000	**0.0105**	**0.0040**
30	0.4690	0.3177	**0.0105**	**0.0131**

B. Experimental Instructions

This is a translation of the original German instructions. Differences in instructions are highlighted.

Figure 4: Instructions

General Instructions for Participants

You are about to take part in an economic experiment. If you read the following instructions carefully, you can earn a substantial amount of money, depending on the decisions you make. It is therefore very important that you read these instructions carefully.

The instructions you have received from us serve your own private information only. **During the experiment, any communication whatsoever is forbidden.** If you have any questions, please ask us. Disobeying this rule will lead to exclusion from the experiment and from any payments.

During the experiment, we speak not of Euro, but instead of Taler. Your entire income is hence initially calculated in Taler. The total number of Taler you earn during the experiment is converted into Euro at the end, at the rate of

1 Taler = 1 Eurocent.

At the end, you will be paid **in cash** the amount of Taler you have earned during the experiment, in addition to 2.5 Euro for taking part.

The experiment is divided into different periods. In total, there are **30 periods**. Participants are divided into groups of five, so your group has another four participants, plus yourself. During these 30 periods, the constellation of your group of five remains unchanged. You are therefore in the same group with the same participants for 30 periods. In each period, you and the other participants in your group will be assigned a random identification number. Please note, however, that this number changes randomly in each round. Group members are therefore **not identifiable** beyond the respective periods. At the beginning, each of the five participants is randomly assigned a role for the duration of the entire experiment. Four participants make decisions in the role of A, and one in the role of B. You keep your role during the entire experiment.

The exact procedure of the experiment is described on the following pages.

Information about the Exact Procedure of the Experiment

Each of the 30 periods has two stages.

Stage 1: Contribution to the Project

Participant A:

At the beginning of each period, each of the four A participants receives an endowment of **20 Taler**. Each A player has to decide how many of the 20 Taler to keep and how many to contribute to the project. All even numbers are possible contributions, i.e., 0, 2, 4, 6, …, 18, 20. All A participants in your group make their respective decisions simultaneously and independently.

After this, the incomes from Stage 1 are calculated:

For each Taler that you keep, you receive exactly one Taler. For each Taler that you and the other participants have invested in the project, each participant receives 0.4 Taler. (Every Taler you invest in the project hence raises the income of each A participant by 0.4 Taler. Conversely, every Taler another participant has invested in the project raises your own income by 0.4 Taler):

Your **income from Stage 1 (Participant A)** is:
+20
– your contribution to the joint project
+ 0.4* total sum of the contributions to the project

The income from the project is calculated by this formula **for all four group members**.

Participant B:

Participants in the B role do not receive any endowment, nor can they contribute anything to the project. Each participant receives 0.4 times the total sum of the contributions to the project. (For every Taler a participant A invests in the project, participant B hence pays 0.4 Taler):

Your **income from Stage 1 (Participant B)** is:
+ 0.4* total sum of the contributions to the project

Please note: Each of the five participants of a group draws the same income from the project, namely 0.4 times the total sum of the contributions to the project, independently of the role they played and of what they invested.

Stage 2: Points Subtracted

(RULES FOR 50 %, ADDITIONAL INSTRUCTIONS IN GREY)

Information

At the beginning of Stage 2, all participants (roles A & B) are informed about the contributions of the (other) A participants to the project. This information (called a "signal") has a 50 % chance of being correct. In other words, in 5 out of 10 cases, the number corresponds to the exact contribution. In the other 5 out of 10 cases, participants see just another random number that does not correspond to the exact contribution. (Any other number apart from the exact contribution has the same chance of appearing; further, all participants receive the same information – except for the person whose contribution is being dealt with – and hence see the same number.)

In stage 2 of every round, all A participants receive an additional 10 Taler. The B participant receives 40 Taler in the second stage of every round.

(RULES FOR DEC)

Distribution of subtraction points

Each participant A can **reduce** the income of other A participants by distributing up to 10 subtraction points. Each of these subtraction points, given by one participant A to another, reduces the latter's income by **3 Taler**. Similarly, each subtraction point distributed costs the distributor **1 Taler and the participant B a further Taler**. You keep all subtraction points that have not been distributed.

Income from the Round

A participant A's income from the round (stages 1 & 2) is therefore:
+ income from stage 1
+ 10 Taler (additional endowment)
– 3 * the sum of the **subtraction points received** from the other A participants
– the sum of the **subtraction points distributed to other A participants** by the A participants

(RULES FOR CEN)

Distribution of subtraction points

By distributing up to 40 subtraction points, participant **B** can **reduce** the income of A participants. Each subtraction point distributed to an A participant by participant B reduces A's income by **3 Taler**. At the same time, each subtraction point distributed costs participant B **1 Taler**, **and every participant A 1/3 Taler (except the one who received the subtraction point)**. B can give a single A participant a maximum of 30 penalty points. You keep all subtraction points that have not been distributed.

Income from the Round

A participant A's income from the round (stages 1 & 2) is therefore:
+ income from stage 1
+ 10 Taler (additional endowment)
– 3 * the sum of the **subtraction points received** from participant B
– 1/3 * the sum of the subtraction points distributed by participant B **to other A participants**

A participant B's income from the round (stages 1 & 2) is therefore:
+ income from stage 1
+ 40 Taler (additional endowment)
– the sum of the subtraction points distributed **by participant B to other A participants**

Information at the End of the Round and Total Income

At the end of each round, you receive from us a detailed overview of your income from the round: Taler you kept for yourself; your income from the project and the resulting income from stage 1; the cost of the subtraction points; the resulting income reduction; the period income.

Your total income at the end of the experiment is the sum of the period incomes.

Is anything still unclear? Please let us know!

Figure 5: Screen Punishment, DEC & CEN