Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Fakultät Wirtschafts- und Sozialwissenschaften

# Information-sensitive Leviathans – the Emergence of Centralized Punishment

**Andreas Nicklisch**
**Kristoffel Grechenig**
**Christian Thöni**

**WiSo-HH Working Paper Series**
Working Paper No. 24
May 2015

WiSo-HH Working Paper Series
Working Paper No. 24
May 2015

# Information-sensitive Leviathans – the Emergence of Centralized Punishment

Andreas Nicklisch, Universität Hamburg
Kristoffel Grechenig, Max Planck Institute for Research on Collective Goods
Christian Thöni, Universität Lausanne

ISSN 2196-8128

Font used: „TheSans UHH“ / LucasFonts



**Kontakt:**
WiSo-Forschungslabor
Von-Melle-Park 5
20146 Hamburg
E-Mail: experiments@wiso.uni-hamburg.de
Web: http://www.wiso.uni-hamburg.de/forschung/forschungslabor/home/

# Information-sensitive Leviathans – the Emergence of Centralized Punishment*

Andreas Nicklisch,[‡] Kristoffel Grechenig,[§]
Christian Thöni[¶]

May 15, 2015

**Abstract**

We study the conditions under which individuals are willing to delegate their sanctioning power to a central authority. We design a public goods game in which players can move between institutional environments, and we vary the observability of others' contributions. Our theoretical analysis suggests that the players choose centralized sanctioning as long as the observability is not too poor. In contrast, our experimental results show that the relative success of centralized sanctioning crucially depends on the interaction between the observability of the cooperation of others and the avoidance of antisocial punishment. While central institutions do not outperform decentralized sanctions under perfect information, large parts of the population are attracted by central institutions that rarely punish antisocially in environments with limited observability.

*Keywords*: centralized sanctions, cooperation, experiment, endogenous institutions
*JEL codes*: C92, D02, H41

*For helpful comments and discussion, we thank Berno Büchel, Ernst Fehr, Guillaume Frechette, Simon Gächter, Sonja Köke, Manfred Milinski, Nikos Nikiforakis, Louis Putterman, Arno Riedl, Bettina Rockenbach, Roberto Weber, and the participants of various seminars at the University of Pennsylvania, the Max Planck Institute (Collective Goods, Bonn), the University of Nottingham, the University of Lausanne, and the University of Cologne. We would also like to thank the Max Planck Society and the German Research Foundation (Ni 1610/1-1) for financial support.

[‡]University of Hamburg and German Research Foundation, Research Unit "Needs Based Justice and Distributive Procedures"

[§]Max Planck Institute for Research on Collective Goods, Bonn

[¶]*Corresponding author:* University of Lausanne, email: christian.thoeni@unil.ch.

# 1 Introduction

Human life in Thomas Hobbes' natural state is lonely, short, and brutal, "a time of war where every man is enemy to every man" (Hobbes, 1651). To redress this grim fate of violence and distrust people appoint a central authority – a *Leviathan* – to enforce cooperative behavior. People voluntarily delegate their sanctioning power to the Leviathan, in the hope for a more efficient outcome.

In contrast to Hobbes' bleak view contemporary research suggests that people successfully use *decentralized sanctions* (peer-to-peer punishment) to enforce cooperation (Ostrom, Walker, & Gardner, 1992; Fehr & Gächter, 2000, 2002) and reach efficient outcomes in the long run (Gächter, Renner, & Sefton, 2008). If human societies are able to organize themselves in a decentralized fashion, one would expect to find many self-governed societies. However, the opposite is the case: we live in a world of mainly centralized sanctions, on the national and even on the supranational level.[1] Why did central authorities emerge in so many modern societies? Under which conditions are people willing to renounce their sanctioning power in favor of a central authority?

To approach these questions we analyze the formation of central authorities theoretically and experimentally. We introduce an environment where players ('citizens') participate in a social dilemma. Prior to this, they can vote by feet for one of three sanctioning institutions: centralized punishment *(CenPun)*, decentralized punishment *(DecPun)*, and a sanction-free institution *(NoPun)*. In *CenPun*, a randomly drawn subject (the 'authority') can punish the citizens in his institution, while citizens are not allowed to punish each other. Authorities participate in the well-being of their citizens, but do not have to bear the costs of punishment themselves. On the other hand, in *DecPun* citizens can punish other citizens in their institution, but bear the costs of punishment themselves.

Our analysis builds on a number of challenges for self governance which have been identified in the literature: antisocial punishment, revenge, and incomplete information. Antisocial punishment (or perverse punishment) refers to the observation that some subjects target their punishment at cooperative subjects. The occurrence of antisocial punishment severly hampers a group's ability to self govern the social dilemma.[2] Related to this issue is the problem of excessive retaliation for received punishment. Here, the findings are

[1] Here, we think of institutions like the United Nations Security Council, or the International Military Tribunal in Nuremberg in 1945/46.

[2] See e.g. Gächter, Herrmann, and Thöni (2005); Bochet, Page, and Putterman (2006); Herrmann, Thöni, and Gächter (2008).

mixed: some authors find that retaliation weakens decentralized punishment (Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008; Nikiforakis, Noussair, & Wilkening, 2012), whereas others do not find such a general effect (Kamei & Putterman, 2013). Finally, it has been shown that decentralized punishment can be inefficient when subjects receive only imperfect information about the contributions of others. Contrary to intuition (but in accordance to the theoretical analysis we develop below) subjects tend to punish more when information becomes more noisily (Grechenig, Nicklisch, & Thöni, 2010; Ambrus & Greiner, 2012).

In our study, we introduce three environments differing with respect to the accuracy of information citizens and the authority receive concerning the contributions of others: in treatment condition ONE, citizens and the authority receive accurate signals about the contributions; in POINT-NINE, citizens and the authority receive signals which are correct in 90 percent of the cases, while in POINT-FIVE, the signals are correct in 50 percent cases. In the two latter treatments punishment decisions have to be taken under incomplete information, introducing the possibility of erroneous punishment acts. Thus, our design allows us to study the choice of punishment institution, i.e., the 'emergence' of *DecPun*, *CenPun*, or *NoPun*, as a function of the information imperfectness induced by our treatment variation.

In our theoretical analysis we show that authorities can implement deterrent sanctioning schemes which are not available under decentralized sanctioning, and attract the entire population unless the signals are very noisy (POINT-FIVE). In contrast, we show experimentally that central authorities emerge only as noise is introduced. Informational imperfection seems to play a crucial role for the centralization of sanctioning. Yet, it is important to stress that not all authorities are successful: only those authorities who mete out less antisocial punishment than the corresponding rate of antisocial punishment in the decentralized sanctioning regime manage to attract a majority of the citizens. Hence, informational complexity in terms of limited observability of others' behavior in combination with the avoidance of antisocial punishment causes the emergence of central punishment.

Our study combines recent discussions on the formation of centralized institutions. Some authors compare the effect of centralized punishment and decentralized punishment on efficiency, showing that regimes with centralized sanctions perform well, even when sanctions are non-deterrent (Tyran & Feld, 2006; Dal Bó, Foster, & Putterman, 2010; Kube & Traxler, 2011), and, in some cases, outperforming decentralized sanctions (O'Gorman, Henrich, & Van Vugt, 2009). Markussen, Putterman, and Tyran (2014) investigate the emergence of centralized sanctions through voting, when centralization is costly. They find that people are particularly responsive to the fixed costs

of having a centralized sanctioning scheme in place, more so than they respond to whether or not the sanctioning scheme is fully deterrent. Putterman, Tyran, and Kamei (2011) allow participants to vote on the sanctioning scheme and find that many groups quickly implement sanctions that induce efficient outcomes. Andreoni and Gee (2012) investigate the formation of centralized sanctions through voting for a sanctioning scheme that punishes only the lowest contributor and find that full contributions are quickly achieved at very low punishment costs. In contrast to our approach, these articles focus on sanctions that are executed automatically; that is, once an implemented rule is violated, players are punished with a certain probability while contribution decisions are perfectly observable.[3]

In contrast, we introduce the authority as a player, who may use punishment in a similar, potentially erroneous or malevolent fashion than his citizens. We do so as we believe that the feature is of particular importance to explain the emergence of authorities in small-scale societies. That is, we compare centralized and decentralized sanctioning when authorities are not equipped with better mechanisms to guide behavior than citizens (e.g., our authorities are not better informed than citizens, nor do they have more efficient punishment technologies than citizens). Rather, our autocratic leader holds absolute and unlimited power within the group; much like in a feudal society he is not appointed by a competitive procedure, but he is merely born into his position.

Following previous works showing that decentralized sanctions prevail over a sanction-free environment (Gürerk, Irlenbusch, & Rockenbach, 2006) and over a pure reputation-building environment (Rockenbach & Milinski, 2006), we let our players choose their institution by leaving societies (exit), but not by votes (voice).[4] As such, we analyze the formation of a central institution as the consequence of an active choice in favor of the authority. That is to say, due to the third alternative *NoPun*, our setting requires citizens to choose actively in favor of one punishment institution allowing us to interpret citizens' institutional choice predominantly as a choice in favor of centralized or decentralized punishment rather than a decision against the alternative sanctioning institution which is not chosen.

Our article is structured as follows. Section 2 describes our basic game

[3]Other related papers include Kosfeld, Okada, and Riedl (2009) and Sutter, Haigner, and Kocher (2010) who endogenize the institutional design to some extent, while they introduce automatically executed centralized punishment as well.

[4]Historically, the importance of exit mechanisms for the organization of tribes, or even the fall of entire nations (e.g. East Germany), is well documented (Hirschman, 1970, 1978). Contemporary exit mechanisms capture competition between authorities for corporations and tax payers, for instance.

and derives theoretical predictions, in section 3 we introduce the experimental setting and discuss behavioral predictions. Section 4 presents the results, and section 5 concludes the paper with a discussion.

# 2 Model

## 2.1 The game

We set up a game which embeds "competition" between centralized punishment, decentralized punishment, and a punishment-free institution in a public goods game. We combine a *voting by feet* mechanism between different sanctioning regimes (Gürerk et al., 2006; Rockenbach & Milinski, 2006) with *imperfect information* about individual contributions (Grechenig et al., 2010; Ambrus & Greiner, 2012). There are ten citizens and one authority. The game consists of three stages. In stage one, each citizen $i$ independently chooses an institution. There are three institutions, each associated with a specific punishment rule: centralized punishment (*CenPun*), decentralized punishment (*DecPun*), and no punishment (*NoPun*). We denote by $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{N}$ the set of citizens in the three institutions. Citizens in a given institution play a public goods game (with punishment) as long as at least two citizens are present.

In stage two, each citizen receives an endowment of 20 experimental currency units (ECU). Citizens simultaneously choose a contribution $g_i \in \{0, 2, 4, ..., 20\}$ to the public good.Each unit contributed to the public good is multiplied by 1.6 and the resulting amount is divided equally among the citizens in the respective institution. This keeps individual payoffs from the public good constant for different group sizes, so that there are no productivity advantages for large groups (Gürerk et al., 2006). A citizen $i$ in the institution *CenPun* earns a profit after stage two of

$$\hat{\pi}_i = 20 - g_i + \frac{1.6}{c} \sum_{k \in \boldsymbol{C}} g_k, \tag{1}$$

where $c \equiv |\boldsymbol{C}|$ denotes the number of citizens in *CenPun*. For citizens in the other two institutions the same payoff function holds with respect to the sets $\boldsymbol{D}$ and $\boldsymbol{N}$.

In stage three, players receive signals about the contribution of the other citizens in their institution. For each citizen $i$ a signal is produced, such that

$$s_i = \begin{cases} g_i & \text{with } prob = \lambda, \\ \widetilde{g}_i & \text{with } prob = 1 - \lambda, \end{cases} \tag{2}$$

where $\widetilde{g}_i$ is randomly drawn from the set $\{0, 2, 4, \ldots, 20\} \setminus \{g_i\}$ with uniform probabilities. Thus, for each citizen, there is an independent random draw determining whether the signal is equal to the true contribution. If not, another independent draw selects a different contribution. The signal $s_i$ is communicated to all other citizens in $i$'s institution, and, in case of *CenPun*, also to the authority.

In addition, all citizens receive an extra endowment of three units. Depending on their institution, players assign punishment points (that is, citizens in *DecPun* and the authority in *CenPun*), and the final payoffs are realized. The three institutions differ only in stage three.

For a citizen in *NoPun* the payoff equals the profit after stage two plus the extra endowment:

$$\pi_i = \hat{\pi}_i + 3 \qquad \forall\, i \in \boldsymbol{N}. \tag{3}$$

In institution *DecPun* all citizens decide simultaneously over punishment $p_{i\to k}$ with $k \in \boldsymbol{D} \setminus \{i\}$. Each punishment point assigned to another citizen leads to a deduction of three units from the punished citizen's payoff and reduces the punisher's payoff by one unit. Each citizen can spend up to her extra endowment for punishment, that is, $\sum_k p_{i\to k} \leq 3$. Units not spent on punishment are credited to the citizens' payoff. For a citizen $i$ in *DecPun*, the payoff equals

$$\pi_i = \hat{\pi}_i + (3 - \sum_{k\in \boldsymbol{D}\setminus\{i\}} p_{i\to k}) - 3 \sum_{k\in \boldsymbol{D}\setminus\{i\}} p_{k\to i} \qquad \forall\, i \in \boldsymbol{D}. \tag{4}$$

In *CenPun* all punishment decisions are delegated to the authority. The authority decides over punishment $p_{\to k}$ with $k \in \boldsymbol{C}$. Like in *DecPun* each punishment point assigned to a citizen leads to a deduction of three units from the punished citizen's payoff and costs one unit. In *CenPun* these costs have to be borne equally by all other citizens in the institution. In sum, the authority can spend up to the extra endowment of all its citizens for punishment, i.e., $\sum_k p_{\to k} \leq 3c$. In addition, maximum punishment targeted at a single citizen is restricted to $3(c-1)$. Units not spent on punishment are credited to the particular citizen's account. Hence, *DecPun* and *CenPun* are identical with regard to financial consequences of punishment. The only difference is that punishment *decisions* are taken by the authority. For citizen $i$ in *CenPun*, the payoff equals

$$\pi_i = \hat{\pi}_i + \left(3 - \frac{\sum_{k\in \boldsymbol{C}\setminus\{i\}} p_{\to k}}{c-1}\right) - 3p_{\to i} \qquad \forall\, i \in \boldsymbol{C}. \tag{5}$$

The authority's payoff equals the average profit after stage two of all citizens in institution *CenPun*

$$\pi_A = \frac{\sum_{i \in \boldsymbol{C}} \hat{\pi}_i}{c} \qquad \text{if } c \geqslant 2. \tag{6}$$

If there is only one citizen in an institution, there is no public good and no punishment. In this case, the citizen receives a payoff of 20. If there are less than two citizens in *CenPun* the authority receives a payoff of 20. All parameters, the signal technology ($\lambda$), and payoff functions are common knowledge.

We vary the information environment $\lambda$ across treatment conditions. In ONE, $\lambda = 1$, citizens and the authority receive perfect information regarding the contributions of members of their institution. In POINT-NINE, $\lambda = .9$, citizens and the authority receive a signal about the others' contributions that displays the accurate information in nine out of ten cases (and a different contribution in the remaining case). Finally, in POINT-FIVE, $\lambda = .5$, players receive accurate information in five out of ten cases.

## 2.2 Theoretical prediction

In the following, we derive predictions concerning punishment, contributions, and institutional choice under standard assumptions. In particular, we assume selfish preferences and risk neutrality. We are interested in equilibria which yield the maximum contributions. We show for ONE and POINT-NINE that there exists an equilibrium where all citizens choose *CenPun* and cooperate fully. In case of POINT-FIVE, the authorities lack the resources to enforce full contributions of all ten citizens. However, there exists an equilibrium in which only two citizens choose *CenPun* and contribute fully.

Using backwards induction, we start by analyzing the punishment stage. The two simple cases are *NoPun* and *DecPun*. In the first there is no punishment, in the second punishment is costly to the punisher, which means that all strategy profiles including punishment acts are not subgame perfect. In *CenPun* things are more interesting. The authority does not bear the cost of punishment. Consequently, the entire set of possible punishment strategies can be part of a subgame-perfect Nash equilibrium in *CenPun*. As the authority's payoff is increasing in its citizens contributions we look for punishment strategies which resolve the social dilemma character of the public goods game in stage two and make it individually rational for the citizens to contribute. That is, we are looking for deterrent punishment strategies which prevent unilateral deviation from contributing a certain level $\bar{g}$ ($20 \geqslant \bar{g} > 0$).

If deterrent punishment is feasible and if it does not require too much antisocial punishment then the game has an equilibrium in which all citizens choose *CenPun* and contribute $\bar{g}$. How could a deterrent punishment strategy look like? If the $c-1$ other citizens contribute $\bar{g}$, citizen $i$'s profit before punishment is

$$\hat{\pi}_i(g_i) = 20 - g_i + \frac{1.6}{c}\Big[(c-1)\bar{g} + g_i\Big]. \tag{7}$$

Taking the derivative with regard to $g_i$ leads to the marginal disutility of contributing, $\frac{1.6-c}{c}$, which is increasing (in absolute terms) in the number of citizens in *CenPun*. To be deterrent a punishment strategy must ensure that the payoff gains of $g_i < \bar{g}$ are set off by an equivalent or larger payoff reduction through punishment. In the following, we focus our attention to the least expensive punishment strategy which exactly matches the profit from every deviation $g_i < \bar{g}$ in expectation. Let $p(s_i)$ be the authority's punishment function, mapping signals into punishment for citizen $i$. If there is no uncertainty ($\lambda = 1$), then a simple linear punishment with the slope $p' = \frac{1.6-c}{3c}$ for all $s_i < 20$ and $p(20) = 0$ suffices to induce full cooperation. With imperfect signals things are slightly more complicated. Using the punishment option inevitably leads to punishment of cooperative subjects. In accordance to the literature we denote this as antisocial punishment hereafter.[5] The value of $\lambda$ determines the informational value of the signal. For $\lambda = \frac{1}{11}$ the signal contains no information about the contribution, which renders deterrent punishment impossible. In the following we restrict our attention to signals with a $\lambda \in \left(\frac{1}{11}, 1\right]$. With such signals the best guess about the true contribution is the signal. Signals of $\bar{g}$ are taken as indication of cooperative behavior and are not punished. Signals above $\bar{g}$ are also not punished. The condition for the least costly punishment function is

$$\begin{aligned}\hat{\pi}_i(\bar{g}) - 3(1-\lambda)\tfrac{1}{10}\sum_{s_i \in \boldsymbol{S}} p(s_i) \\ = \hat{\pi}_i(g_i) - 3\lambda p(g_i) - 3(1-\lambda)\tfrac{1}{10}\left[\sum_{s_i \in \boldsymbol{S}} p(s_i) - p(g_i)\right],\end{aligned} \tag{8}$$

where the left-hand side shows expected payoff of contributing $\bar{g}$, consisting of the stage two payoff minus three times the expected antisocial punishment points. Antisocial punishment occurs with probability $(1-\lambda)$ and

[5] In our context, we define *antisocial punishment* as punishment targeted towards citizens who contribute equal or more than $\bar{g}$, irrespective of the signal the punisher gets. Punishment towards a citizen with a *signal* weakly larger than $\bar{g}$ (which does not make sense from a deterrence perspective) is referred to as *misguided punishment*. For $\lambda = 1$ the two definitions are equivalent, but not for $\lambda < 1$. When analyzing the data we cannot observe $\bar{g}$ and we will use the mean contribution (or the mean signal) instead.

consists of the expected punishment for all possible wrong signals, where $\boldsymbol{S} = \{0, 2, 4, \ldots, 20\}$ is the set of all signals. The right-hand side shows the expected utility for any contribution $g_i < \bar{g}$, consisting of the deviation payoff from stage two minus the 'correct' punishment, as well as the punishment triggered by false signals. Here, we have to subtract the punishment for the true contribution $g_i$ from the sum (this term is zero on the left-hand side). Rearranging leads to

$$\hat{\pi}_i(g_i) - \hat{\pi}_i(\bar{g}) = 3\lambda p(g_i) - \tfrac{3}{10}(1-\lambda)p(g_i), \tag{9}$$

where the left-hand side shows the increase in stage two profits from deviating and the right-hand side shows the increase in expected punishment from deviating. The latter consist of the punishment based on the true signal reduced by the decrease in punishment due to false signals. In case of perfect signals the latter would be zero, in case of uninformative signals ($\lambda = \frac{1}{11}$) the right-hand side equals zero, which confirms our statement above that deterrence is impossible under these circumstances. Using equation (7), we can solve equation (9) for the punishment function dependent on the signal[6]:

$$p^*(s) = \left[\frac{(10c-16)(\bar{g}-s)}{3c(11\lambda-1)}\right]_{[0}. \tag{10}$$

For $s < \bar{g}$ the least costly deterrent punishment is linear in the signal $s$ (we omit the subscript, because the punishment strategy is the same for all citizens). Furthermore it is decreasing in $\lambda$ and increasing in $c$ and $\bar{g}$, that is, more noisy signals, larger groups, and higher contributions require stronger punishment.[7]

Having shown that deterrent punishment is possible for $\lambda > \frac{1}{11}$ raises the question of its feasibility. Given our design of the punishment mechanism the authority faces two 'incentive compatibility constraints'. The first one ($IC_t$) is due to the restriction on total punishment, the second one ($IC_i$) by the restriction on individual punishment. To derive $IC_t$ we calculate the expected punishment expenditures necessary to deter a group of citizens with one deviator. We take the case of the most expensive deviation, which is a

[6]The notation $[a]_{[0}$ is equivalent to $\max\{0, a\}$.

[7]We derived $p^*(s)$ under the assumption that each citizen is punished only dependent on his own signal. Alternatively, the authority could adopt even more complicated punishment strategies $p(\boldsymbol{s})$, where $\boldsymbol{s} = (s_i, s_j, \ldots)$ is a vector of all signals observed. We also derived punishment strategies for the rules (i) punish only the citizen(s) with the lowest signal(s) in $\boldsymbol{s}$, (ii) punish the lowest signal only if unique, and (iii) punish if and only if there is a single signal lower than $\bar{g}$. The expected expenditures for disciplining a fully cooperative group are identical to the case of $p^*(s)$ for all punishment strategies (i), (ii) and (iii).

contribution of zero. We relate the expected expenditures to the authorities budget constraint, which is $3c$:

$$\lambda p^*(0) + (1-\lambda)\tfrac{1}{10} \sum_{s \in \boldsymbol{S}\setminus\{0\}} p^*(s) + (c-1)\,(1-\lambda)\tfrac{1}{10} \sum_{s \in \boldsymbol{S}} p^*(s) \leqslant 3c. \qquad (11)$$

The first two elements of the left-hand side refer to the expected punishment for the free-rider, followed by the expected punishment for the remaining citizens who contribute $\bar{g}$. This expression allows us to find the enforceable contribution levels dependent on the number of citizens and the noise in the signals.

For $IC_i$ recall that maximum punishment imposed on a single citizen is $3(c-1)$. Depending on $\bar{g}$, $\lambda$, and $c$ there are situations in which this constraint does not allow for the punishment necessary to deter free riding, that is, $p^*(0) > 3(c-1)$. Finally, in addition to these two incentive compatibility constraints ($IC_t$, $IC_i$), we also have to satisfy a participation constraint, ensuring that the punishment costs of a cooperative group do not surpass the efficiency gains created by contributing $\bar{g}$ instead of zero in another institution. Similar to the expression in equation (11) we calculate the expected punishment costs for a fully cooperative group of citizens. Different from before, we have to take into account that the income reduction is not only due to received punishment, but also due to the financing of the punishment of others in the group, that is, we have to ensure that

$$4(1-\lambda)\tfrac{1}{10} \sum_{s \in \boldsymbol{S}} p^*(s) \leqslant \tfrac{3}{5}\bar{g}. \qquad (12)$$

Figure 1 shows the numerical results for these three conditions. All lines indicate combinations of $\lambda$ and $c$ for which one of the conditions holds with equality and ruling out the cases to the left of the line. Solid (long dashed, short dashed) lines indicate the incentive compatibility constraints for $\bar{g} = 20, (18, 16)$. The monotonically increasing lines refer to the constraint on total punishment, whereas the mostly decreasing lines indicate the constraints on maximum punishment for a single citizen. Dotted lines indicate the participation constraints, and bold lines indicate the envelope of all constraints. Typically one of the two incentive compatibility constraints is binding, the participation constraint is only binding for $\bar{g} = 16$ and for $c \geqslant 7$.[8]

[8]For the more complicated punishment strategies $p(\boldsymbol{s})$ discussed in footnote 7 the expected punishment costs as calculated in equation 11 tend to be smaller, thus relaxing $IC_t$. However, this does not open up more equilibria for $\lambda = .5$, because $IC_i$ is violated

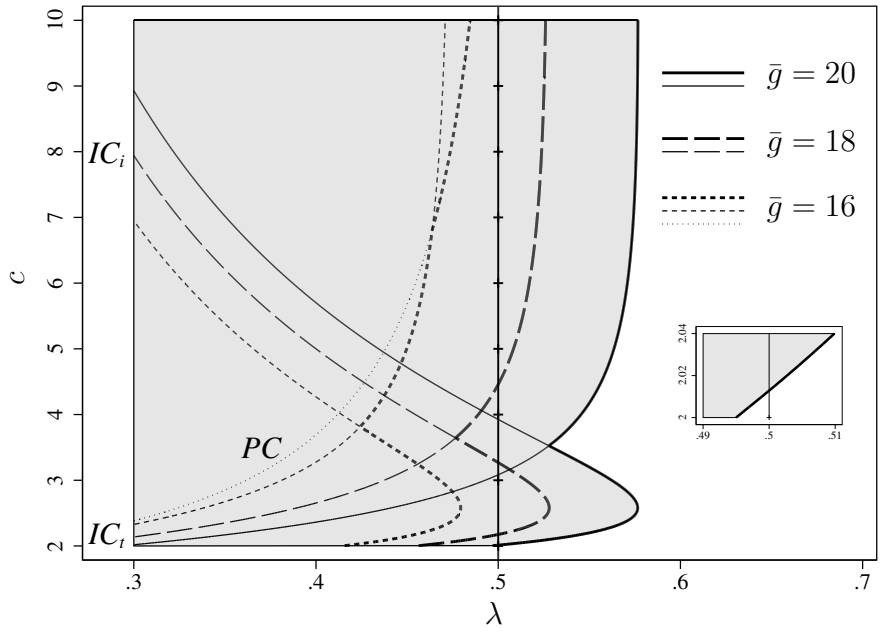


Figure 1: Feasibility of cooperative outcomes in *CenPun*, dependent on the quality of the signals $\lambda$ and the number of citizens $c$. Positively sloped dashed or solid lines indicate the incentive compatibility with respect to total punishment ($IC_t$), mostly downwards sloped lines show the incentive compatibility for individual punishment ($IC_i$). In both cases the area to the right is feasible. The dotted line indicates the participation constraint for $\bar{g} = 16$ (not binding for higher $\bar{g}$). Bold lines indicate the envelope of all constraints. The small subfigure shows the region around $\lambda = .5$ for small groups enlarged.

For $\lambda > .58$ none of the constraints are binding and fully cooperative outcomes can be enforced by the authority. The grey area indicates constellations for which a fully cooperative outcome cannot be enforced. For $\lambda = .5$ (which we implemented in the experiment) the constraint on total punishment is binding in case of $c > 3$, while $c = 3$ is ruled out by the constraint $IC_i$, leaving $c = 2$ as the only possibility (see small subgraph enlarging the area). Going for $\bar{g} = 18$ relaxes the constraints (long dashed line) such that, in addition to $c = 2$, $c = 4$ becomes feasible. Enforcing $\bar{g} = 16$ (or lower) is

for all cases with $c > 2$. Intuitively, these strategies use punishment less often, but require stronger punishment when applied.

feasible for all numbers of citizens.

This allows us to formulate a set of Nash equilibrium strategies for the three levels of $\lambda$ used in our experiment. For $\lambda = .9$ and 1 there exists an equilibrium in which all citizens choose *CenPun* and contribute fully. Equilibrium strategies are as follows: The authority punishes all citizens in $\boldsymbol{C}$ based on the signals according to $p^*(s)$ as defined in equation (10).

All citizens play the same equilibrium strategies:

- Stage 3: No punishment in *DecPun*: $p_{i \to k} = 0 \; \forall \; k \in \boldsymbol{D}$
- Stage 2: Contribute $g_i = 0$ if in *DecPun* or *NoPun*, contribute $g_i = 20$ if in *CenPun*
- Stage 1: Choose *CenPun*

For $\lambda = .5$ this equilibrium is not feasible. The authority's preferred outcome would be to attract only two citizens, because then the fully cooperative outcome is enforceable. To do so the authority might play a punishment strategy with punishment according to $p^*(s)$ in case of $c = 2$ and punishment independent of the signal otherwise. The best response of the citizens would be two out of ten entering, while the others choose a different institution. The problem is that this punishment strategy is not subgame perfect. If, for example, four citizens choose *CenPun* the authority would prefer to implement a punishment inducing contributions of 18. Thus, for any punishment strategy it must hold that it enforces the highest level of contributions given $c$. Despite this additional condition the authority can reach maximum payoff in equilibrium by playing the following strategy in $\lambda = .5$:

- Punish according to $p^*(s)$ for the highest $\bar{g}$ feasible given $c$
- In addition, use the remaining punishment points to punish all citizens in $\boldsymbol{C}$ except citizen $i$ and $j$ by equal amounts, independent of the signal.

This strategy ensures that the participation constraint is only met for citizen $i$ and $j$ if deterrent punishment leaves enough resources to punish citizens other than $i$ and $j$. For our parameters this is the case. For instance, in case of $c = 4$ enforcing $\bar{g} = 18$ requires 8 out of 12 punishment points and the citizens earn an expected payoff of 25.8 when punished according to $p^*(s)$. Using the remaining 4 punishment points to reduce the income of two citizens reduces their expected income by 6 units each, making them worse off than the outside option of 23. Consequently, the citizens' strategies for $\lambda = .5$ are

- Stage 3: No punishment in *DecPun*: $p_{i \to k} = 0 \; \forall \; k \in \boldsymbol{D}$

- Stage 2: Contribute $g_i = 0$ if in *DecPun* or *NoPun*. In *CenPun* contribute the highest enforceable contribution given $c$, that is, $g_i = 20$ if $c = 2$, $g_i = 18$ if $c \leq 4$, and $g_i = 16$ else.
- Stage 1: Citizen $i$ and $j$ choose *CenPun*. All other choose *NoPun*.

To conclude, under standard assumptions we expect central punishment to emerge in case of $\lambda \geq .9$, implementing full contribution by all citizens and no (ONE) or moderate antisocial punishment (POINT-NINE). For $\lambda = .5$, only a subgroup of citizens choose *CenPun*, while the others choose *NoPun* (or *DecPun*); contributions in *CenPun* are high (i.e., $g_i \geq 16$), but the enforcement requires substantial antisocial punishment.

# 3 Experimental Game and Behavioral Predictions

The experiment is played in matching groups of eleven subjects. Prior to the start of the game we randomly allocate one subject in each matching group to the role of the authority and ten subjects to the role of the citizen. Roles remain the same throughout the experiment.

Because the game is fairly complicated, and because we think that interesting things might unfold with time we implement a repeated game of 32 periods. Participants know that they play the game for the finite number of periods.[9] Since we want to provide the three institutions with some time to establish cooperation before they are put into competition with other institutions, the citizens in our experiment choose their institution every fourth period only. Thus we implement a game with eight phases consisting of four periods each. At the beginning of each phase all subjects allocated to the role of citizens choose one of the three institutions and remain there during the phase.

Each period consists of three steps, a contribution step, a punishment step, and an information step. Appendix A.2 shows the information provided on the screens during the experiment. In the punishment step, all citizens and the authority receive the signals from the citizens in their institution. If applicable, citizens or the authority choose their punishment points. The identification number of citizens are randomly reassigned between periods.

[9]English translations of the instructions are reported in Appendix A.1. Before the experiment starts, subjects have to solve a set of control questions on the computer screen.

In the information step, citizens learn their period payoff including received punishment.[10]

At the beginning of each phase all citizens are informed about the outcome in all institutions (see screenshot in Figure A1). In particular, when choosing an institution citizens know (i) the number of citizens (ii) the average contribution, and (iii) the average profit in all three institutions and for all previous periods. At this point the information is undistorted.

Inspired by our theoretical analysis of the underlying game in the previous section, we expect that the central authority attracts all citizens as long as there is no or moderate noise in the signals, while if there is substantial noise the central authority attracts only a small fraction of citizens.

Our stark theoretical results rest on the assumption that citizens do not punish in *DecPun* due to the positive marginal costs of punishment. A large body of evidence on public goods games with punishment shows that this is not a good description of actual behavior: many individuals are willing to use costly punishment (Chaudhuri, 2011). If subjects can choose between decentralized punishment and a no punishment institution the majority eventually ends up in the punishment institution under perfect information (Gürerk et al., 2006). In the light of these stylized facts from previous experiments, we expect that citizens manage to reach and maintain high contributions in *DecPun* in ONE. In such an environment, it is difficult for the central authority to offer an advantage. In addition, there are potential reasons against centralized punishment: citizens may fear that the central authority might excessively punish due to the fact that it is not costly to him, or they may have a simple preference to retain their punishment power (Fehr, Herz, & Wilkening, 2013). We thus expect citizens to prefer *DecPun*, so that central punishment rarely emerges in ONE.

What changes under imperfect information? The evidence on experiments with decentralized punishment suggests that noise hampers the functioning of peer punishment, but does not necessarily discourage the citizens from using the punishment option (Grechenig et al., 2010). Punishing in a noisy environment bears the risk unintentionally meeting out antisocial punishment, that is, punishment targeted towards cooperative citizens. Evidence from experiments with perfect information suggests that the occurrence of antisocial punishment strongly reduces cooperation and motivates retaliatory counter punishment (Herrmann et al., 2008). While the central authority cannot avoid antisocial punishment as well, we expect a competitive advantage for *CenPun* under imperfect information because it prevents counter

[10]Citizens receive no information about their own signal, that is, they do not know whether other players are correctly informed about their contribution or not.

punishment. For this reason, we expect the emergence of central punishment in the two noisy environments POINT-NINE and POINT-FIVE.

# 4 Results

We run a series of laboratory experiments with a total of 15 sessions with 30 independent populations (330 participants, 110 per treatment). Each subject participated in only one population. The experiments were conducted at the laboratory for economic experiments (EconLab) at the University of Bonn with mostly undergraduate students from various fields. Six percent of participants were non-students, 56 percent of participants were females, and age ranged between 18 and 64 (median 22). The experiment was programmed in z-Tree (Fischbacher, 2007); we used ORSEE (Greiner, 2004) for recruiting. A session lasted for about 120 minutes. Payoffs were converted at an exchange rate of 1 Euro per 75 ECUs; payoffs accrue over all periods. Subjects earned on average 15.64 Euros, including a show-up fee of 4 Euros.

## 4.1 Choice of institutions

For the choice of institution in the first phase, *NoPun* attracts the majority of the population in all treatments. About two thirds of the subjects choose this institution in POINT-NINE and even more in the other two treatments. This is in line with the results of Gürerk et al. (2006), who also find that their punishment institution is not popular early in the game. Centralized punishment initially attracts 21 percent of the citizens in POINT-NINE, compared to 13 and 7 percent in ONE and POINT-FIVE, respectively. Albeit moderate in size, these differences in the initial choice of institutions are significant across treatments ($p = .027$, Fisher's exact test). Over time, most citizens move to the two punishment institutions. Panel A of Figure 2 shows the average choice of institutions across all periods. In ONE, the modal choice is *DecPun*. In POINT-NINE, the modal choice is *CenPun*, although by a small margin over the two other institutions. More uncertainty seems to favor *NoPun*, which is the modal choice in POINT-FIVE. The distribution of institution choices is significantly different across treatments ($F(2.49, 72.26) = 3.41$, $p = .029$ for all periods; $F(2.85, 82.65) = 3.14$, $p = .032$ for the final phase, Pearson $\chi^2$ statistic with correction for dependence within group, see Rao and Scott, 1984).

Panel B of Figure 2 shows the average profits earned in the three institutions over time. Profits tend to be higher for the two institutions allowing for punishment, but overall differences are not pronounced. This is not surpris-

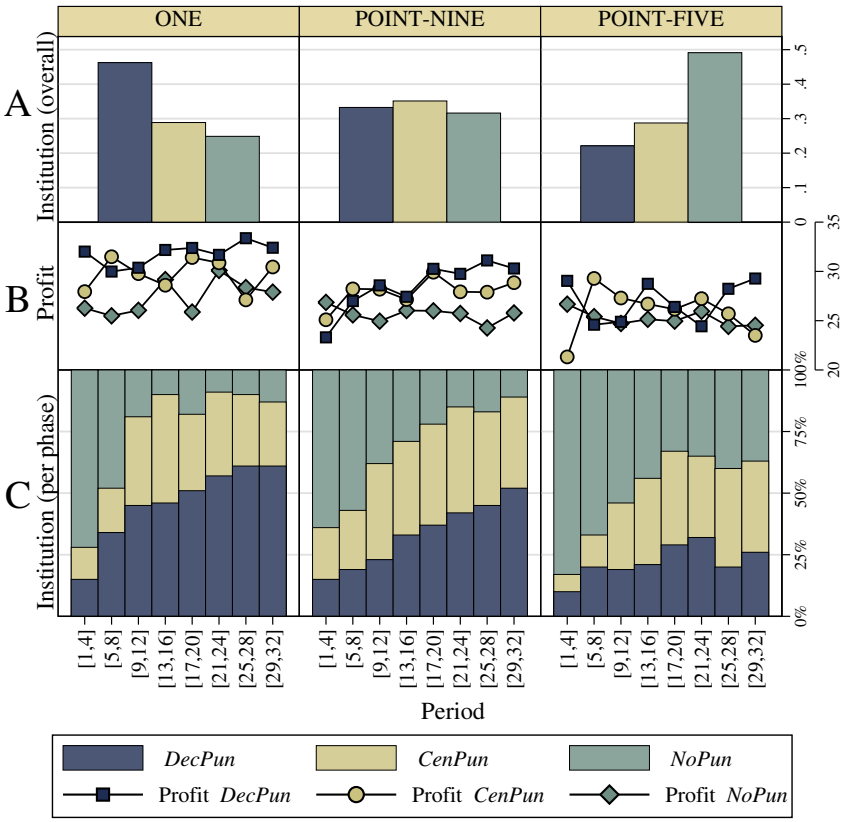


Figure 2: Panel A: Average choice of institution over all periods and by treatment. Panel B: Average profits in *NoPun*, *DecPun*, and *CenPun* across time. Dots show averages in a phase of four periods. Panel C: Choice of institution during the eight phases.

ing, given that there is free movement between the institutions every fourth round. A comparison across treatment shows that profits are decreasing in the noise level of the signals. Average profits drop from 29.2 in ONE to 27.3 in POINT-NINE, and 25.4 in POINT-FIVE. The differences are highly significant ($p = .002$, Kruskal-Wallis test on matching group averages).

Panel C of Figure 2 shows the relative share of the institutions over time. In all treatments, *NoPun* loses a lot of citizens during the first three phases. Most of the adjustments happen through the first half of the 32 periods and we observe relatively stable shares of institutions in the second half of the experiment in ONE and POINT-FIVE. In POINT-NINE, the share of *CenPun* is stable, but *NoPun* loses in favor of *DecPun* throughout the 32 periods.

When citizens can move between institutions, they are informed about the outcomes in the three institutions. In particular, citizens learn (i) the number of citizens, (ii) the average contribution, and (iii) the average profits earned in each of the three institutions in all previous periods. We use multinomial probit models to explain the choice of institution between phases. For each citizen we observe seven institution choices with information about the outcome of the prior phase. In Model (1) of Table 1 we explain the choice of institution by the average profit of the citizens in each institution in the previous phase.[11] We use two dummies for the treatments ONE and POINT-NINE, with POINT-FIVE being the omitted case. We also add two dummies for the institution in which the subject is currently in, with *NoPun* as the omitted case, and we add a linear time trend (variable *Phase*). The treatment dummies indicate that citizens are less likely to choose *NoPun* over *DecPun* in the two treatments with relatively accurate or perfect information. There is strong inertia in the institution choice, that is, having been in *NoPun* before significantly increases the chance of choosing *NoPun* relative to *DecPun*, as shown by the significant negative effects of both institution dummies. The coefficients of the three profit variables show that this information is indeed a strong determinant for the institution choice. Observing high profits in *NoPun* significantly increases the probability of choosing *NoPun* over *DecPun* for the next phase, while the opposite is true for high profits in *DecPun*. The profits in *CenPun* do not seem to affect the choice between *NoPun* and *DecPun*. The estimates for choosing *CenPun* (the second set of covariates in Table 1) show a very similar pattern. High profits in *CenPun* increase the probability of choosing that institution for the next phase over *DecPun*, while the opposite is true for high profits in *DecPun*.

While the relation between relative profits and institution choice is strong, it is not informative with regard to the ultimate causes of the relative attractiveness of the institutions, because profits are merely a result of the activities in a given phase. The profits are mainly linked to contributions (for *NoPun* they are linearly dependent). If we replace the profits by contributions in Model (1) of Table 1 we get very similar results (not shown in the table), that is, high contributions in an institution increase the probability of choosing the respective institution. However, the main source of the relative success of the two punishment institutions should be determined in the way the citizens and the authority use the punishment option.

[11] In case there were no citizens in a given institution we cannot observe a profit. In the estimates we use the same profit as in the case when there is only one citizen in a given institution.

Table 1: Choice of institution.

| | Dependent variable: Institution in $t+1$ | | | |
|---|---|---|---|---|
| | (1) | | (2) | |
| Choose *NoPun* | | | | |
| ONE | −0.417*** | (0.139) | −1.001*** | (0.240) |
| POINT-NINE | −0.298** | (0.131) | −0.644*** | (0.205) |
| *DecPun* | −1.740*** | (0.181) | −2.137*** | (0.196) |
| *CenPun* | −0.824*** | (0.163) | −0.894*** | (0.188) |
| Phase | −0.003 | (0.026) | −0.113*** | (0.031) |
| Profit *NoPun* | 0.100*** | (0.018) | | |
| Profit *DecPun* | −0.124*** | (0.014) | | |
| Profit *CenPun* | −0.007 | (0.011) | | |
| Free-rider pun × *DecPun* | | | 0.045** | (0.020) |
| Antisocial pun × *DecPun* | | | 0.117*** | (0.045) |
| Free-rider pun × *CenPun* | | | 0.074** | (0.035) |
| Antisocial pun × *CenPun* | | | 0.040 | (0.066) |
| Constant | 1.753*** | (0.586) | 1.607*** | (0.162) |
| Choose *CenPun* | | | | |
| ONE | −0.208 | (0.165) | −0.473* | (0.287) |
| POINT-NINE | −0.072 | (0.162) | −0.183 | (0.285) |
| *DecPun* | −0.778*** | (0.190) | −1.286*** | (0.180) |
| *CenPun* | 0.787*** | (0.162) | 1.219*** | (0.168) |
| Phase | −0.006 | (0.028) | −0.034 | (0.032) |
| Profit *NoPun* | 0.035** | (0.016) | | |
| Profit *DecPun* | −0.123*** | (0.015) | | |
| Profit *CenPun* | 0.133*** | (0.015) | | |
| Free-rider pun × *DecPun* | | | 0.028 | (0.021) |
| Antisocial pun × *DecPun* | | | 0.114** | (0.046) |
| Free-rider pun × *CenPun* | | | −0.010 | (0.028) |
| Antisocial pun × *CenPun* | | | −0.125** | (0.057) |
| Constant | −1.005* | (0.577) | 0.313 | (0.202) |
| Wald $\chi^2$-test | 1724.2 | | 606.2 | |
| $p$ | 0.000 | | 0.000 | |
| $N$ | 2100 | | 2100 | |

*Notes:* Multinomial probit estimates. Dependent variable: Chosen institution for the next phase (*DecPun* is the omitted case). Independent variables are treatment dummies (POINT-FIVE as omitted case), dummies for the institution in the previous phase (*NoPun* as omitted case), Phase, average profits in the actual phase in the respective institution, and Free-rider and Antisocial punishment in the respective institutions during the previous phase. Robust standard errors, clustered on matching group, in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In Model (2) of Table 1 we investigate the use of the punishment option as a determinant of institution choice. Just adding the frequency or strength of punishment used in a given institution is, however, not an adequate measure of how well cooperation norms are enforced. Punishment can not only be targeted at low contributors, but also at high contributors. We classify received punishment into free-rider punishment (if the punished citizen contributed less than the group average) or antisocial punishment (otherwise). We replace the covariates for the profits by variables measuring free-rider, and antisocial punishment, interacted with the dummy for the two institutions allowing for punishment. The results in the upper half of the table show that punishment in *DecPun* increases the probability of leaving the institution in favor of *NoPun*. Interestingly this holds both for free-rider punishment and antisocial punishment, although the latter effect seems to be much stronger. In the lower half of the table we find clear evidence that the occurrence of antisocial punishment is decisive in the choice between the two institutions allowing for punishment. High antisocial punishment in *DecPun* significantly increases the probability of choosing *CenPun*, and vice versa. The use of free-rider punishment, on the other hand, does not significantly affect the choice between these two institutions.

## 4.2 Heterogeneous Leviathans

Given that antisocial punishment is a crucial determinant of entry into and exit out of an institution we now investigate whether individual heterogeneity in punishment behavior of authorities and citizens explain the relative success of the punishment institution. We use the measure of antisocial punishment to determine the 'quality' of the central authority. In particular, we calculate for each population the average received antisocial punishment of the citizens in *DecPun* and *CenPun*. Populations in which we observe weaker antisocial punishment in *CenPun* than in *DecPun* are classified as populations with a 'good' authority. Conversely, if the authority metes out more antisocial punishment than the citizens we speak of a population with a 'bad' authority. Our classification is based on the data of phases 1-7 and we explain the institutional choices in the final phase (8). In POINT-NINE this criterion leads to an equal split of the matching groups, while in POINT-FIVE and ONE we classify 60 percent of the matching groups as populations with a good authority.[12]

[12]In ONE we have three matching groups for which the punishment data is missing because the citizens never chose the respective institution. In these cases we use the average of the corresponding figures in the other groups in the same treatment as an estimate for antisocial punishment.

Panel A of Figure 3 shows that authorities attract only a negligible fraction of the citizens in ONE and POINT-NINE when they mete out a lot of antisocial punishment relative to the citizens in *DecPun*. Panel B shows that good authorities manage to attract a larger share than bad authorities in all treatments. However, only under imperfect information *CenPun* is clearly the modal choice. Under perfect information not even good authorities are able to gain the support of the majority of the population.

Panels C and D of Figure 3 provide information about the stability of the population in *CenPun*. Bars show the fraction of citizens in this institution, divided into incumbents (darker part) and immigrants (lighter part). Incumbents are citizens who were already in *CenPun* in the preceding phase; immigrants are citizens who were previously in *DecPun* or *NoPun*. The graph shows that bad authorities have a high turnover: most of the time, more than half of their population are immigrants. Populations of good authorities are much more stable, with a large fraction of the citizens remaining in the institution.

Instead of dividing the observations in two groups we can also use the difference between the antisocial punishment in *DecPun* and of the authority as a continuous measure of an authority's relative performance in punishing. If we use OLS to regress the share of the population in *CenPun* in the final phase (the middle bars in Figure 3) on this measure we observe a highly significant positive effect for POINT-NINE ($\beta = .414, p = .010$, robust standard errors, group averages as observations), but not for the other two treatments.

Our result that authorities who do not mete out antisocial punishment are able to attract a majority of the citizens in POINT-NINE but not in ONE is stable to alternative specifications of good and bad authorities. A first alternative specification is to look only at antisocial punishment in *CenPun* (ignoring punishment in *DecPun*) and to perform a median split according to this measure (phase 1-7). The left panel of Figure 4 shows the resulting institution choice in the last phase for the good authorities according to this specification. A second alternative is to focus on assigned instead of received punishment. In the treatments with imperfect information this distinction is important, because some authorities might be classified as bad authorities even if they never punish citizens for which they receive an above average signal. In this specification we calculate a measure for *misguided punishment*, i.e., punishment targeted at citizens with at or above average *signals*. We use the observations from phase 1-7 for each authority and perform a median split to identify the good authorities. The right panel of Figure 4 shows the institutional outcome. In case of POINT-NINE the distribution of institutions in the final phase is almost identical for all specifications of good authorities. In ONE we confirm the result that authorities do not manage to attract a

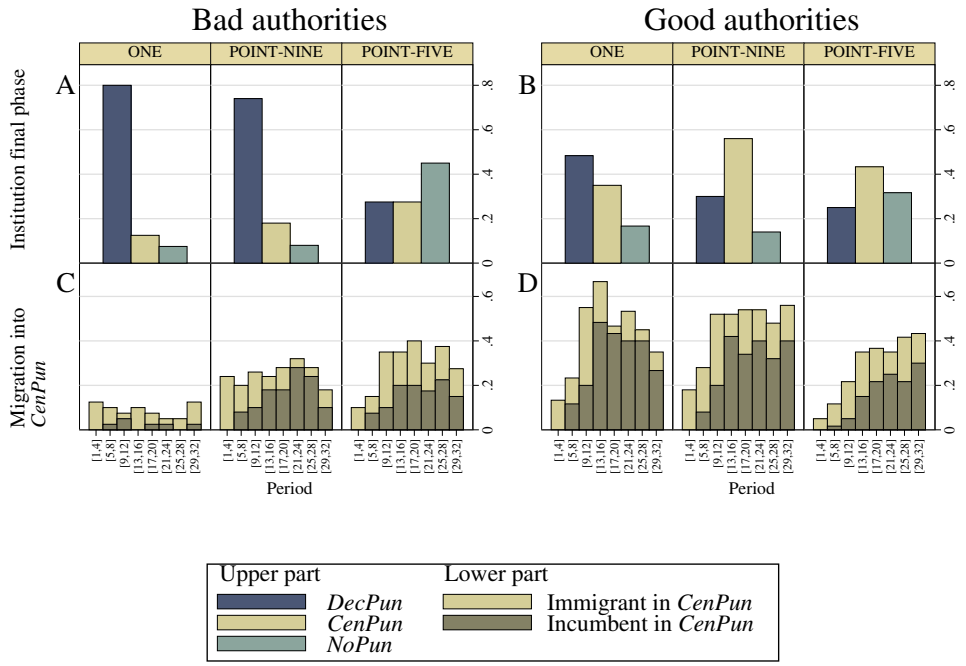


Figure 3: Choice of institution in the final phase of the game for matching groups with bad (panel A) and good authorities (panel B). Bars show the fraction of participants choosing *DecPun, CenPun, NoPun*, separated by the treatments with perfect information ONE and imperfect information POINT-NINE, POINT-FIVE. Panels C and D: Migration patterns in *CenPun*. The dark part of the bars shows the citizens who were in *CenPun* already in the previous phase (incumbents); the light part shows the immigrants.

majority of the population, while the results in POINT-FIVE seem to be more volatile. Interestingly, authorities who avoid misguided punishment seem to be able to attract a large share of the population.

## 4.3 Punishment strategies

In section 2.2 we derived a formal expression for the minimum deterrent punishment (equation 10). We can now compare the authorities' punishment decisions with this theoretical benchmark. The expression requires to specify a contribution level $\bar{g}$. For the two treatments in which enforcing full

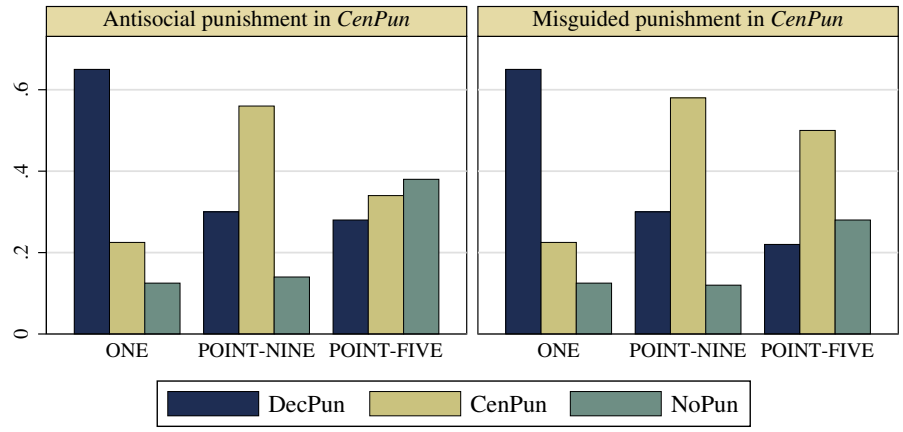


Figure 4: Institution choices for populations with a good authority for two alternative specifications of good authorities. Left panel: Median split of all authorities according the antisocial punishment (punishment of at or above average contributions). Right panel: Median split of the authorities according to misguided punishment (punishment of at or above average signals).

contributions is feasible (ONE and POINT-NINE) the authorities might want to set $\bar{g} = 20$ and consequently punish all citizens with signals below the maximum according to equation 10 or stronger. In groups which are not fully cooperative this strategy would lead to a lot of antisocial punishment which might drive the citizens out of the institution when they get the chance to leave. A more reasonable punishment strategy might thus be to abstain from punishing the relatively cooperative citizens and set $\bar{g}$ to a 'typical' contribution level in the given group and period. In the following analysis we set $\bar{g}$ equal to the median signal an authority receives in a given period.[13] Using equation (10) we calculate the predicted punishments for all constellations of signals authorities face, and compare them to the actual punishment decisions.

The left panel of Figure 4.3 shows the results for the three treatments. On the vertical axis we depict the difference between the signal and the median signal in the group $(s_i - \bar{g})$. For example, a value of $-20$ refers to the case where the signals indicate that the citizen is a free rider and half (or

[13]For even numbers of signals we slightly deviate from the usual calculation of the median and take the higher of the two middle values, i.e., in case of the signals $\{20, 18, 12, 0\}$ we set $\bar{g} = 18$.

more) of the other citizens contribute fully. The bars indicate the average number of punishment points meted out for the respective deviation. The horizontal lines show the average predicted minimal deterrent punishment, strongly decreasing in the signal for free riders (negative deviations) and zero thereafter.[14] The top left panel of Figure 4.3 shows the results for ONE. Punishment clearly follows the theoretical pattern, but tends to be lower than predicted, with the exception of moderate negative deviations of four to two units. In addition, there is very little misguided punishment. For POINT-NINE we observe a similar result, where the punishment for negative deviations tend to be very close to deterrent punishment or slightly above. Again there is basically no misguided punishment (due to false signals there is still antisocial punishment). In POINT-FIVE we observe a totally different pattern. For negative deviations punishment is much lower than deterrent punishment. In addition, punishment seems almost invariant across the deviation classes.[15]

In a next step we want to contrast the authorities' use of the punishment to the punishment of the citizen's in *DecPun*. For the theoretical benchmark we assume that each citizen calculates $\bar{g}$ on the basis of the signals she receives, but including her own contribution. In addition, for groups with more than two citizens we assume that each citizen punishes other citizens by $\frac{1}{n-1}$ of the minimal deterrent punishment according to equation 10. The right panel of Figure 4.3 shows the results for *DecPun*. In ONE we observe that punishment for negative deviations is substantially higher than predicted. This holds also for POINT-NINE, where in addition we observe a clear increase in misguided punishment relative to *CenPun*. Finally, POINT-FIVE leads again to punishments far from deterrent and largely invariant in the deviation.

Taken together these observations suggest a rationale for the shift of the competitive advantage from *DecPun* towards *CenPun* once we move from perfect information to moderate noise. Contrary to the notion of second order free-rider problems in the punishment stage, it seems that decentralized norm enforcement is excessive relative to minimal deterrent punishment. In ONE, however, this might not render *DecPun* unattractive, because there is

[14]The expression in equation 10 is linearly decreasing in $s_i$ for $\bar{g} > s_i$, but the horizontal lines in Figure 4.3 are not. The reason for this is that we combine all cases with various values for $s_i - \bar{g}$ and $c$ into a single average per bar.

[15]In section 2.2 we have shown that the punishment budget does not allow to enforce full contributions in POINT-FIVE. A possible explanation for the fact that punishments are not as high as predicted might be that authorities hit the constraints in total punishment. The data shows that this is not the case, as less than two percent of the authorities' punishment decisions in POINT-FIVE exhaust the budget (in the other two treatments the number is even lower).

a simple strategy to avoid excessive punishment, namely contributing. This is no longer the case in POINT-NINE. Also here citizens tend to punish low signals stronger than necessary, which sometimes leads to antisocial punishment, due to wrong signals. Since citizens cannot avoid this punishment, an institution which does not mete out more punishment than necessary to deter free riding might get the competitive edge. In addition it seems also that misguided punishment is a lot more frequent among citizens in *DecPun* than among authorities.

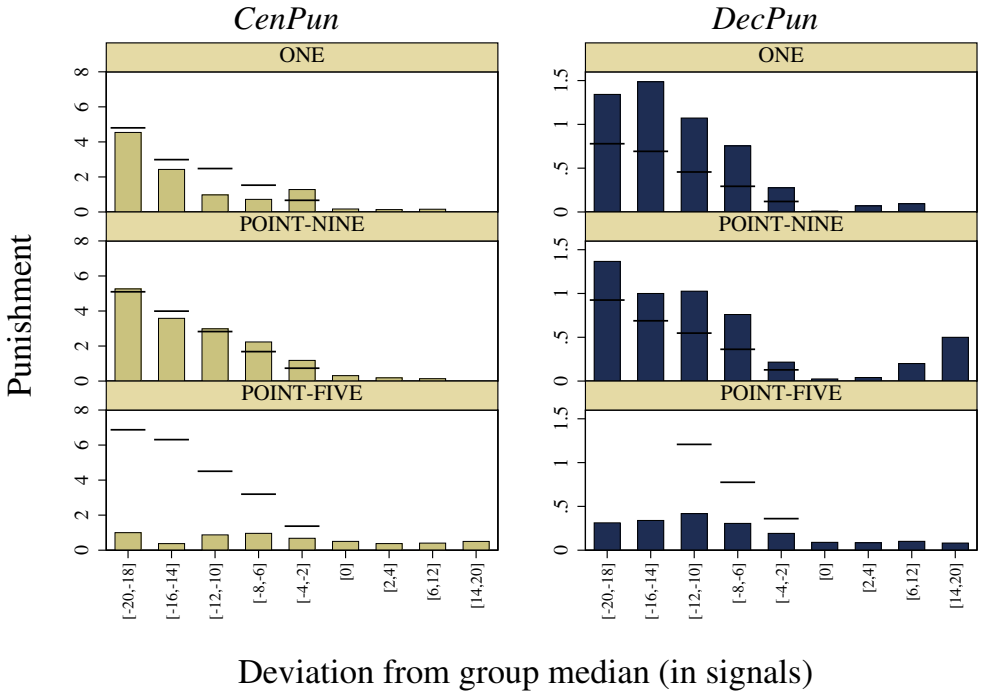


Figure 5: Predicted and actual punishment of authorities in *CenPun* (left panel) and citizens in *DecPun* (right panel) for the three treatments. Bars show average punishment targeted at a citizen dependent on the difference between the citizen's signal and the median signal in the group. Horizontal lines show the average of all deterrent punishments according to the theoretical prediction (values for the two first bars in the bottom right panel are outside the plotted range).

# 5 Discussion

Our study analyzes the formation of central authorities. The experimental results show that centralized punishment institutions – despite the perfect alignment of group interest and the interest of the authority – emerge only under two conditions: imperfect monitoring and the availability of a central authority who refrains from punishing cooperative citizens. We consider our treatment variations as prototypical for various epochs of the evolution of social structures in humans. Small-scale societies allowing for nearly perfect observation of others tend to apply decentralized punishment regimes. In larger societies with increasing agglomeration and complexity, it becomes increasingly difficult to monitor others' behavior. These are the circumstances, in which people are willing to sacrifice some of their autonomy and delegate the sanctioning power to a Leviathan.

The reason for this pattern are competitive advantages of centralized punishment in comparison to peer punishment under noise: it seems that decentralized punishment leads to excessive sanctioning. Yet, in ONE, citizens can easily avoid excessive punishment by contributing (nearly) fully, while citizens cannot avoid the excessive and/or antisocial punishment in POINT-NINE. Here, a central punishment institution, which does not mete out more punishment than necessary, can get the competitive edge and attract large shares of the population.

Applied to modern western societies, we might underestimate the attractiveness of centralized punishment in our experiment, as there are arguably better selection mechanisms for authorities and institutional restraints against antisocial punishment in place. Presumably, these societies come close to the outcome of good authorities in POINT-NINE, where *CenPun* is clearly the dominant institution. In times of social unrest and destabilized law enforcement systems, however, punishment by authorities becomes more erratic. Under these circumstances centralized sanctions lose their competitive advantage and, if possible, citizens migrate to other institutional arrangements.

Recently, the appearance of new media like social networks and mobile communication technologies give rise to another interesting development: it leads to increasing transparency of actions within groups. As a consequence, we might expect a decentralization of the societal structures. The latest developments on the administration of mass protests during the Arab Spring via social networks are an example for this development (Hussain & Howard, 2013). Whether this is a first indication for a general shift towards more decentralized organizational structures is too early to tell.

# References


Ambrus, A. & Greiner, B. (2012). Imperfect public monitoring with costly punishment: an experimental study. *American Economic Review*, *102*(7), 3317–3332. doi:10.1257/aer.102.7.3317

Andreoni, J. & Gee, L. K. (2012). Gun for hire: delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics*, *96*(11-12), 1036–1046. doi:10.1016/j.jpubeco.2012.08.003

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, *60*(1), 11–26. doi:10.1016/j.jebo.2003.06.006

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, *14*(1), 47–83. doi:10.1007/s10683-010-9257-1

Dal Bó, P., Foster, A., & Putterman, L. (2010). Institutions and behavior: experimental evidence on the effects of democracy. *American Economic Review*, *100*(5), 2205–2229. doi:10.1257/aer.100.5.2205

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, *33*(1), 145–167. doi:10.1007/s00199-007-0212-0

Fehr, E. & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994. doi:10.1257/aer.90.4.980

Fehr, E. & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140. doi:10.1038/415137a

Fehr, E., Herz, H., & Wilkening, T. (2013). The lure of authority: motivation and incentive effects of power. *American Economic Review*, *103*(4), 1325–1359. doi:10.1257/aer.103.4.1325

Fischbacher, U. (2007). Z-tree: zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Gächter, S., Herrmann, B., & Thöni, C. (2005). Cross-cultural differences in norm enforcement. *Behavioral and Brain Sciences*, *28*(6), 822–823. doi:10.1017/S0140525X05290143

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 2008. doi:10.1126/science.1164744

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment despite reasonable doubt - a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies*, *7*(4), 847–867. doi:10.1111/j.1740-1461.2010.01197.x

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

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*(5770), 108–111. doi:10.1126/science.1123633

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. doi:10.1126/science.1153808

Hirschman, A. O. (1970). *Exit, voice and loyalty. Responses to decline in firms, organizations and states*. Cambridge MA: Harvard University Press.

Hirschman, A. O. (1978). Exit, voice, and the state. *World Politics*, *31*(1), 90–107. doi:10.2307/2009968

Hobbes, T. (1651). *The Leviathan*. Oxford World's Classics Series. Oxford University Press, Incorporated, 1996.

Hussain, M. M. & Howard, P. N. (2013). What best explains successful protest cascades? ICTs and the fuzzy causes of the Arab Spring. *International Studies Review*, *15*(1), 48–66. doi:10.1111/misr.12020

Kamei, K. & Putterman, L. (2013). *In broad daylight: fuller information and higher-order punishment opportunities can promote cooperation*, Working Paper 2012-3, Brown University, Department of Economics.

Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, *99*(4), 1335–1355. doi:10.1257/aer.99.4.1335

Kube, S. & Traxler, C. (2011). The interaction of legal and social norm enforcement. *Journal of Public Economic Theory*, *13*(2006), 639–660. doi:10.1111/j.1467-9779.2011.01515.x

Markussen, T., Putterman, L., & Tyran, J.-R. (2014). Self-organization for collective action: an experimental study of voting on sanction regimes. *Review of Economic Studies*, *81*(1), 301–324. doi:10.1093/restud/rdt022

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics*, *92*, 91–112. doi:10.1016/j.jpubeco.2007.04.008

Nikiforakis, N., Noussair, C. N., & Wilkening, T. (2012). Normative conflict and feuds: the limits of self-enforcement. *Journal of Public Economics*, *96*(9-10), 797–807. doi:10.1016/j.jpubeco.2012.05.014

O'Gorman, R., Henrich, J., & Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B*, *276*(1655), 323–9. doi:10.1098/rspb.2008.1082

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review*, *86*(2), 404–417. doi:10.2307/1964229

Putterman, L., Tyran, J.-R., & Kamei, K. (2011). Public goods and voting on formal sanction schemes. *Journal of Public Economics*, *95*(9-10), 1213–1222. doi:10.1016/j.jpubeco.2011.05.001

Rao, J. N. K. & Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, *12*(1), 46–60.

Rockenbach, B. & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, *444*(7120), 718–23. doi:10.1038/nature05229

Sutter, M., Haigner, S., & Kocher, M. G. (2010). Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies*, *77*(4), 1540–1566. doi:10.1111/j.1467-937X.2010.00608.x

Tyran, J.-R. & Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics*, *108*(1), 135–156. doi:10.1111/j.1467-9442.2006.00444.x

# A Appendix

## A.1 Experimental instructions

This section includes a translation of the instructions handed out on paper (original instructions were in German). The instructions are identical for all treatments and all roles (citizen, authority), with exception of the description of signal accuracy which we put in brackets.

### 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 do not speak of Euro, but 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

**75 Taler = 1 Euro.**

At the end of the experiment, you will be paid **in cash** the amount of Taler you have earned during the experiment, in addition to 4 Euro for taking part in the experiment.

The experiment is divided into different rounds. In each round, you will be given an identification number, so that your decisions in the course of a round can be attributed to you. Please note that, after each round, the identification number allocated to you and the other members of your group changes randomly. Group members therefore cannot be identified beyond the rounds. All decisions are made **anonymously**, i.e., none of the other participants is told the identity of a person who made a particular decision. The payoff is also anonymous, i.e., no participant is told how high another participant's payoff is.

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

# Information about the Exact Procedure of the Experiment

**General Information**

At the beginning of the experiment, you are randomly assigned to one of two halves, each of which has 11 participants. During the entire experiment, you interact only with participants from your half. At the beginning, one of the 11 participants is chosen at random for the entire duration of the experiment, receiving a different task from the one which the other participants are assigned to

**Procedure**

The experiment consists of 32 rounds. At the very beginning and, from then on, every four rounds (i.e., in rounds 1, 5, 9, 13,..., 29), you may choose a group. There are three different groups: A, B, and C. Each of the 32 rounds consists of 2 stages. In the first stage, you choose a contribution to the joint project. In the second stage, you can influence the income of the other participants in your group by means of subtracting points. Groups A, B, and C differ in this regard.

**Group Choice**

Every participant chooses a group:

| | **Influencing the Income of the other Group Members:** |
|---|---|
| Group | A: No points subtracted |
| | B: Mutual point subtraction |
| | C: Point subtraction by the extra participant |

You will find more details below about the way participants can influence the income.

**Stage 1: Contribution to the Joint Project**

In each round, you receive an endowment of $20$ Taler. It is up to you to decide how many of the 20 Taler you wish to contribute to the joint project. All even numbers are possible contributions, i.e., $0, 2, 4, 6, \ldots, 18, 20$. All other participants in your group make the same decisions simultaneously. After this, the incomes from Stage 1 are calculated:

| Your **income from Stage 1** is: |
|---|
| 20 $-$ your contribution to the joint project |
| $+$ 1.6 $\times$ the average contribution |

You therefore keep all Taler that you have not contributed to the project. In addition, you receive 1.6 times the average of the contributions from all group members (the average of the contributions is the sum of the contributions from all group members to the project, divided by the number of group members).

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

**Please note:** Each group member receives the same income from the project, regardless of how much he or she has paid in, i.e., each group member profits from all contributions to the joint project.

**Stage 2: Points Subtracted**

**(i) General Information**

In Stage 2, all other participants in your group (A, B, or C) and the extra participant (if you are in group C) are told their contribution (henceforth referred to as the signal). [This signal is correct with a probability of 50% (90%). In other words, in 5 (9) out of 10 cases, the figure that the other participants see in your group corresponds to your actual contribution. In the remaining 5 (9) out of 10 cases, the other participants see a random other number that does not correspond to your contribution (here, all numbers can appear with equal probability).] You also receive a signal for each of the other members of your group, as well as for their contributions. [This information is also correct with a probability of 50% (90%).] In addition, you receive 3 extra Taler in Stage 2 of each round.

**(ii) Groups**

**Group A: No Point Subtracted**

If you have chosen Group A, then you cannot take any action during this stage.

Your **income from Stage 2** is therefore:
3

**Group B: Mutual Points Subtraction**

If you have chosen Group B, then you may **reduce** or **leave unchanged** the income of the other members of your group. You must decide how many of the **3 Taler** you wish to spend on **distributing** subtraction points to other group members. Every subtraction point that you give to another group member reduces this member's income by **3 Taler**. (Similarly, your own income is reduced by 3 Taler per subtraction point distributed by another group member to you.) At the same time, every subtraction point distributed by you to others costs you **1 Taler**. You **keep** the remaining Taler.

Your **income from Stage 2** is therefore:
3 − the sum of the subtraction points you distribute to other group members in Group B
− 3 × the sum of the subtraction points you receive from other participants in Group B

**Group C: Point Subtraction by the Extra Participant**

In Group C, the **extra participant**, rather than the group members, decides on the distribution of subtraction points (see the passage "Extra Participant (Group C)"). The extra participant also receives the information on the contribution decisions of the Group C participants. [This information, too, has a 50% (90%) likelihood of being correct.] If the extra participant gives

you subtraction points, then your income is reduced by **3 Taler**. The cost of subtraction points that the extra participant gives to another Group C participant must be evenly divided among all other Group C participants. For instance, if 5 participants are in Group C and the extra participant gives one participant 2 subtraction points, then the remaining four participants each have to shoulder the cost of 0.5 Taler.

Your **income from Stage 2** is therefore:
3 − (the sum of the subtraction points from the extra participant to others) / (Number of participants in Group C - 1)
− 3 × the sum of the subtraction points you receive

**Extra Participant (Group C)**

Should you have become the extra participant, the following refers to you. Unlike the other participants, you do not decide between the groups, and you cannot choose a contribution to the joint project either. However, like the other participants, you receive a signal about each Group C player's contribution. [This signal is correct with a probability of 50% (90%). In other words, in 5 (9) out of 10 cases, the figure corresponds to the actual contribution of the respective group members. In the remaining 5 (9) out of 10 cases, you see a random other number that does not correspond to your contribution (here, all numbers can appear with equal probability).]

Your task is to choose the subtraction points for Group C. You may give each individual participant in Group C separate subtraction points. In total, you can distribute a maximum number of subtraction points that corresponds to three times the number of Group C participants.

Your income is determined by the mean income of Group C participants in Stage 1 (i.e., prior to the income reduction caused by subtraction points). The higher the contributions in Group C are, the higher your income is as an extra participant.

Your **total income** in this round is therefore:
Average income of Group C participants in Stage 1

**Special Case: Only Group Member**

Should you be the **only member** in a group (in Group C, apart from the extra participant), you receive 20 Taler in Stage 1 and no Taler in Stage 2, i.e., your income for the round is 20 Taler. You have no possibility to take action" neither at the first nor at the second stage. If you are an extra participant or if there are zero participants or only one participant in Group C, you also receive **20 Taler** and have no possibility to take action.

**Information at the End of the Round**

At the end of the round, you receive a detailed overview of the results of your group. Each group member is told the own contribution to the project, the income from Stage 1, subtraction points distributed (if possible), subtraction points received (if possible), the income from Stage 2, and the income from the round. Every four rounds, you may choose whether you would like to be in Group A, B, or C for the next four rounds. For this decision, you are given an

overview of the average round incomes of the last four rounds in Groups A, B, and C.

**Round Income and Total Income**

Your income from Stage 1 plus your income from Stage 2 taken together generate your income in each round. The total income from the experiment is calculated by adding the incomes from all 32 rounds.

Is anything unclear? Please contact someone in charge of heading the experiment!

## A.2 On screen parts

After reading the instructions, the subjects had to solve six control questions on screen. The questions included hypothetical combinations of contribution and punishment decisions and the participants had to calculate the resulting payoffs. After all participants completed the control questions the experiment began. Only then did participants learn whether they were assigned the role of a citizen or an authority. Figure A1 shows the screen for the institution choice in period 5. Each subject is informed about the number of subjects, average contributions and payoffs in all three groups at the time of the decision about the institution for the next phase of four periods. In the punishment stage authorities were presented with the contributions of the participants in their group and had to choose deduction points. Figure A2 shows an example of a screen for stage 2. The screen for the citizens in *DecPun* looks alike with the exception that one of the rows contains the subjects own contribution and does not take an input on deduction points. Likewise, the screen for the citizens in *CenPun* as well as in *NoPun* contains the identical information about the contributions but does not take input.

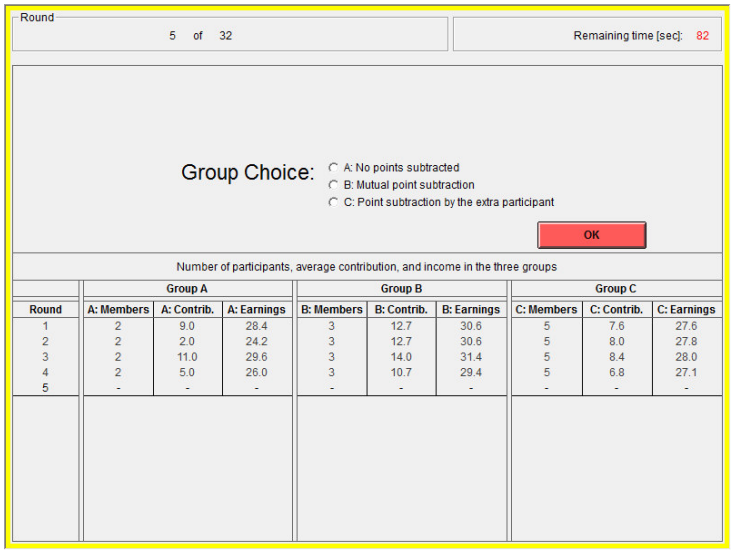


Figure A1: Screen of institution choice stage after the first phase of four periods.

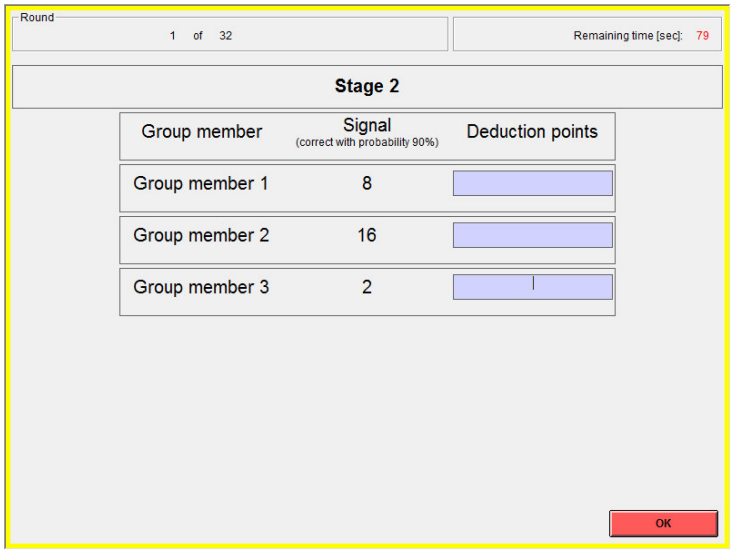


Figure A2: Screen of punishment stage for a central authority with three subjects in POINT-NINE.