Endogenous emergence of institutions to sustain cooperation*

Ernst Fehr and Tony Williams‡
University of Zurich

November 14, 2013

‡JOB MARKET PAPER

Abstract

Formal and informal institutions, such as laws and social norms, are pervasive in daily life. They help maintain cooperation by coordinating and constraining individuals' behaviors. However, our understanding of the comparative benefits and the endogenous emergence of institutions remains limited. Here, we study the emergence and performance of sanctioning institutions in a public goods context when individuals are free to migrate between different institutions. We show experimentally that efficient peer and centralized sanctioning emerge as dominant institutions that immediately generate and maintain high levels of cooperation without much need for costly punishment. The quick establishment of high cooperation is due to both the self-selection of prosocial individuals into these institutions and the institutions' intrinsically beneficial properties. In addition, voluntary migration into the centralized sanctioning institution leads to the selection of stable prosocial leaders who refrain from antisocial punishment, while remnants of antisocial punishment still exist under peer punishment.

Keywords: cooperation, punishment, endogenous institutions, public goods

JEL Classification Numbers: D02, D03, D72, H41

*Fehr and Williams: Department of Economics and Laboratory for Social and Neural Systems Research, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zürich (email: ernst.fehr@econ.uzh.ch, tony.williams@econ.uzh.ch). We thank Sam Bowles, Rob Boyd, Joe Henrich, and Pete Richerson for comments regarding the anthropology literature on small-scale societies. We are also thank Björn Bartling, Donja Darai, Holger Herz, and Frédéric Schneider for comments on an earlier draft and and Roberto Weber for comments on several drafts. We also thank participants at the 2013 Economic Science Association World Meetings and the Workshop on Norms and Cooperation, both in Zürich, for helpful comments. Williams especially thanks Dan Burghart for helpful discussions and Tim Salmon for comments, discussions, and a great deal of encouragement. Williams gratefully acknowledges funding from the Swiss National Science Foundation under SNF Doctoral Program (ProDoc) Grant PDFMP1-123113/1.

Institutions are pervasive in social and economic life. They "are the humanly devised constraints that shape human interaction" which include both formal institutions such as laws and constitutions as well as informal institutions such as social norms, conventions, and taboos (North, 1990). Institutions shape economic and social incentives and are therefore of paramount importance for the economic performance of individuals, groups, companies and, perhaps, even countries. In the long run, however, institutions are themselves subject to individual and political choices, and may thus be viewed as an equilibrium outcome in a broader "game" in which different institutions compete for the support of the population.

In this paper we study the endogenous emergence of a particularly important "humanly devised constraint" - sanctioning institutions - in the context of public goods provision. Throughout human evolution, social groups faced important public goods problems that ranged from the provision of social insurance through food sharing among hunter-gatherers and cooperation during warfare between neighboring groups to the provision of effort among coworkers who receive a group bonus in case of high profits. The experimental literature on public goods provision (e.g. Fischbacher et al., 2001) has shown that many people are willing to contribute voluntarily to public goods if others do so as well, but it is generally not possible to sustain a high level of cooperation if free-riders face no sanctions (Ostrom et al., 1992; Fehr and Gächter, 2000). Historically, peer sanctions are probably the oldest form of sanctioning that emerged among hunter-gatherers long before humans developed more centralized sanctioning institutions that involved judges and central enforcement of punishments. However, peer punishment has been shown to generate high initial costs because of coordination failure among peers and because a considerable amount of initial sanctioning is necessary to establish the credibility of the punishment threat and, in small groups, punished individuals may not necessarily respond to sanctioning in a prosocial manner (Gächter et al., 2008; Dreber et al., 2008). Peer punishment is, in particular, often associated with "antisocial punishment," i.e. when low contributors punish individuals who make above average contributions to the public good (Gächter et al., 2008).

The short and medium run inefficiency of uncoordinated peer punishment raises the question whether and how human groups are capable of avoiding the high initial costs of peer punishment. We study this question in an experimental environment in which individuals are free to sort themselves into different sanctioning institutions. Ethnographic evidence indicates that early human groups were characterized by high mobility and frequent migration in and out of existing groups (Boehm et al., 1996; Kaplan et al., 2005; Wiessner, 2005; Mathew and Boyd, 2011). Thus, allowing individuals to leave and join groups freely seems to capture an important component of social life in the early evolu-

tion of humans.[1] In our experiment, individuals can sort into four different institutions; within an institution, they can contribute to, and benefit from, a public good that only benefits members of the institution. For simplicity, and to have a stark contrast between individual and collective interest, it is in an individual's rational self-interest to contribute nothing to the public good when he or she faces no sanctions for free-riding, but group welfare is maximized if everybody contributes the whole endowment to the public good.

One of the available institutions is characterized by the absence of any explicit opportunity for the sanctioning of individual free-riders ("no punishment"). The second institution provides an opportunity for each group member to sanction any other group member after they have observed each of their contributions to the public good. We denote this institution as "uncoordinated peer punishment" because it does not offer any explicit possibility to coordinate the group members' contribution or punishment activities. This institution has dominated the experimental economics literature in recent years starting with Ostrom et al. (1992) and Fehr and Gächter (2000). We add a cheap-talk normative request to peer punishment in a third institution, denoted by "coordinated peer punishment," as minor communication could act as a coordination device for contributions, determine when and who should be punished, and potentially affect beliefs about others' preferences. Here, each institution member can state how much he or she thinks everyone in the institution should contribute. The average of these statements is then communicated to every institution member before the contribution decision. The rationale for the coordinated peer punishment institution is that sanctioning typically does not take place in a normative vacuum. Rather, people often sanction for a reason, i.e., they punish what they consider as normatively inappropriate behavior. It thus makes sense to allow them to express their normative views and provide them with feedback about the average view in the group. The fourth and final institution ("coordinated central punishment") maintains the cheap-talk normative request but allows for the delegation of punishment to a single (central) authority elected by the group while also socializing the cost of punishment. This type of institution is prevalent in both small-scale societies (e.g. village elders and tribal chiefs, with collective punishment by the group) and large-scale societies (e.g. police, courts, and prisons funded by taxes).

Our results show that both coordinated peer punishment and centralized punishment function very well and establish extremely high cooperation levels *from the beginning with little need for sanctions*. After an initial adjustment phase, subjects thus predominantly choose these two institutions, while the other two institutions - no punishment and uncoordinated peer punishment - become depopulated. In fact, the uncoordinated

[1]Our set up is also related to Tiebout (1956) who argues that public goods are largely provided at the level of the local community and that consumer-voters will "vote with the feet" by moving to communities that best satisfy their preferences.

peer punishment institution is almost never chosen, even at the very beginning.

The centralized punishment institution completely removes the inefficiencies of uncoordinated peer punishment and already leads to payoff levels that are significantly greater than in "no punishment" in the first period. In addition, centralized punishment removes antisocial punishment. The high efficiency of this institution is based on the two key facts. First, many prosocial individuals (i.e., those with prosocial other-regarding preferences) enter this institution at the very beginning, leading to high normative contribution requests and the selection of a prosocial central authority. Rather than being merely cheap talk, the high normative requests are associated with high actual contributions - individuals seem to use the average contribution request as a coordination device. Subjects thus quickly establish a strong cooperative culture in the centralized punishment institution. The second reason for the superiority of centralized punishment is due to its intrinsically beneficial properties - even in the absence of endogenous sorting of subjects, this institution is capable of producing high cooperation with comparably little punishment costs.

Coordinated peer punishment shares many of the good properties of centralized punishment. Many prosocial individuals immediately enter this institution; they establish very high normative requests followed by equally high contributions. However, this institution requires more actual sanctions during the first few periods, and some antisocial punishment still persists. Therefore, initial payoffs are not larger than in "no punishment" but - in contrast to the uncoordinated punishment institution (Gächter et al., 2008) - payoffs are never smaller than in "no punishment." In fact, coordinated peer punishment outperforms "no punishment" in terms of overall payoff after only three or four periods. Taken together, our results show that efficient punishment institutions emerge endogenously through a competitive process in an environment in which people can "vote with their feet." Prosocial individuals play a key role in this process because they quickly establish a cooperative culture that considerably shortens the length of time that it takes to render an institution efficient. While uncoordinated peer punishment incurs large initial costs, the combination of endogenous sorting of prosocial individuals with the possibility of coordinating group behavior through normative requests very quickly makes both peer punishment and centralized punishment the superior institutions.

Our results speak to a growing body of research on endogenous choice and cooperation. Broadly speaking, these papers fall into three categories: endogenous groups with fixed institutions (e.g. Ahn et al., 2008, 2009), fixed groups with endogenous institutions (e.g. Kosfeld et al., 2009; Sutter et al., 2010), and endogenous groups with endogenous institutions (e.g. Gürerk et al., 2006). The last category is the smallest but also best captures the idea of "voting with feet" to select communities that satisfy the individual's

preferences due to Tiebout (1956). Our paper falls into this category.

Gürerk et al. (2006, 2011, 2013) precede our work in allowing for both endogenous groups and endogenous institutions. Their treatments restrict subjects to only two institutions, (i) a non-sanctioning institution and (ii) an uncoordinated peer-sanctioning institution. Their treatments allow for punishment, reward, or both in the peer-sanctioning institution. However, only one sanctioning institution is available in any treatment and subjects cannot express their normative requests in any of their treatments. Our study shows that the existence of these requests is not innocuous because - if available - subjects immediately leave the uncoordinated punishment institution in favor of coordinated peer punishment or centralized punishment. In addition, we provide insights into the key role of prosocial individuals for the quick and smooth functioning of punishment institutions because we also elicit an independent measure of subjects' prosociality.

Like us, Grechenig et al. (2013) extend the punishment institutions to also allow for centralized punishment. In contrast, however, our central authority is elected by institution members each period and also bears part of the cost of punishment, while they exogenously and permanently assign subjects to be the central authority. The election allows us to examine to whom subjects delegate authority, while the exogenous and permanent assignment in Grechenig et al. (2013) rules out this possibility. We also allow for a cheap-talk normative request that serves as a coordination device for contributions, which helps to quickly establish a cooperative culture. Finally, our work differs from Grechenig et al. (2013) because they neither provide an independent measure of subjects' prosociality nor do they compare the functioning of institutions under endogenous sorting and under exogenous assignment of individuals to institutions.

Our approach may also be a useful complement to research on the persistence of macro-institutions and historical development. Lab experiments are not a substitute for empirical data and identification of effects due to natural experiments and the use of instruments. However, laboratory experiments can provide a controlled setting to test theories that emerge from naturally-occurring data without the need for (non-experimenter) exogenous variation. Nunn (forthcoming) includes a discussion of mechanisms underlying the persistent effects of institutions in historical development and focuses on culture, norms, genetics, and coevolutionary processes. Much of the current knowledge of these various factors has been shaped by work involving cross-cultural lab-in-field experiments (Henrich et al., 2006, 2010; Marlowe et al., 2008).[2] Kimbrough et al. (2008) is a note-

[2]For example, an unresolved question in the macro-institutions literature that is also relevant for development policy is whether institutions will continue to be successful when exogenously imposed in new environments. Acemoglu et al. (2011) find evidence supportive of exogenous imposition in the French Revolution, while Berkowitz et al. (2003a,b) argue that the evidence for exogenously imposed institutions following World War II and the fall of the Soviet Union has been much more mixed. Lab and field experiments can potentially identify the important common features of institutions and which

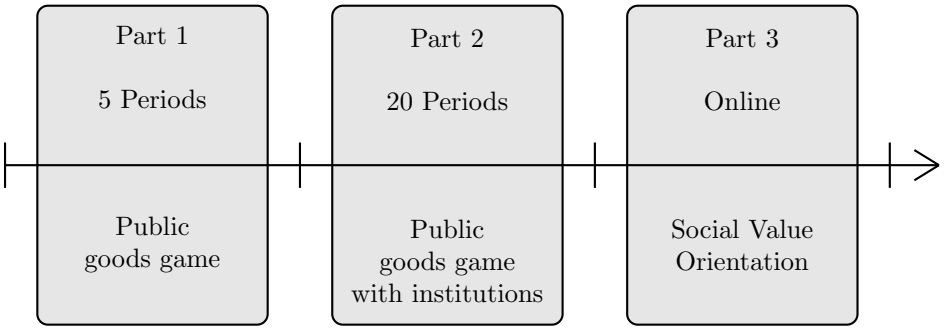


Figure 1: Experiment timeline.

worthy example of how controlled lab experiments can help inform our understanding of historical development, in which they focus on the emergence of long-distance trade.

The remainder of the paper is organized as follows. Section 1 presents our experimental design in detail. Section 2 presents our results. Section 3 concludes the paper and discusses open questions and possible fertile future studies.

1 Experimental design and procedures

The experiment consists of three parts. Parts 1 and 2 are conducted in the lab during the same session. Part 3 is conducted online after subjects leave the lab and provides an independent measure of subjects' social preferences.

Subjects are initially assigned to a large group of size $N \in \{9, 11, 12\}$ that stays fixed for Parts 1 and 2 of the experiment. We attempted to have 12 members in all groups but occasionally used smaller groups due to subjects registering for the study but failing to show up to the lab. Subjects are randomly assigned a unique identification number from the set {1,2,...,N} which also stays fixed for Parts 1 and 2 of the experiment.

Part 1 consists of five periods of a typical public goods game without punishment. Part 2 consists of 20 periods of a public goods game in which subjects can endogenously form subgroups by adopting different punishment institutions; we also include a control treatment to disentangle institutional effects from selection effects (see Section 1.2.4). Part 3 is conducted online and measures social preferences using the Social Value Orientation scale of Murphy et al. (2011). The experiment timeline is summarized in Figure 1.

specific features may be effective in similar contexts (eg. common culture) but likely to fail in alternative environments. They may also be able to do so at lower cost while improving the effectiveness of costly large-scale interventions by identifying critical aspects in advance.

Parameter	Value	Meaning
e	20	Endowment
m	1.5	Multiplier for contribution to public good
N	9, 11, or 12	Group size

Table 1: Parameter values used in Part 1.

1.1 Part 1: Public goods game

Subjects begin by playing five rounds of a typical public goods game without punishment. At the beginning of each period, subjects receive an endowment of points, e, and can contribute any amount $g_i \in \{0, 1, \ldots, e\}$ to a group project. Each point contributed to the group project is multiplied by m and shared equally among all N group members. Each point not contributed to the project goes into a private account. Thus, per-period earnings are given by

$$\pi_i = e - g_i + \left(\frac{m}{N}\right) \sum_{j=1}^{N} g_j. \tag{1}$$

A social dilemma exists whenever (i) $\partial \pi_i / \partial g_i = (m/N) - 1 < 0$ and (ii) $\partial (\sum_{j=1}^{N} \pi_j) / \partial g_i = m - 1 > 0$ for all $g_i > 0$. Condition (i) means that own payoff is decreasing in own contribution to the project, so that free-riding is individually optimal for payoff-maximizers. Condition (ii) means that aggregate payoffs are increasing in own contribution to the project, so that a full contribution by every group members is socially optimal.

The marginal per-capita return (MPCR) is given by (m/N) and is decreasing in group size N. We set $m = 1.5$ in the experiment, so that MPCR ranged from 0.125 when $N = 12$ to 0.167 when $N = 9$. Previous experiments suggest that these values for the MPCR and group size should lead to the breakdown of cooperation within five periods.[3] Subjects should therefore directly experience the public goods problem during Part 1 of the experiment and have a potential motivation to form institutions to establish and maintain cooperation in order to improve their own payoffs. In each period and after making private contribution decisions, group members are informed of every group member's contribution. In addition to $m = 1.5$, we set $e = 20$ and $m = 1.5$. The parameter values are summarized in Table 1.

[3]Hamman et al. (2011) used an MPCR of 0.15 with fixed group size of 9. Average contributions began around 45% of the endowment in the first period and declined to about 15% in period 5.

1.2 Part 2: Public goods game with endogenous punishment institutions

Subjects remain in the same group and retain the same identification number in Part 2 of the experiment. Part 2 lasts for 20 periods and provides subjects with the opportunity to form institutions with other group members after experiencing the public goods problem during Part 1. At the beginning of each period, subjects individually select into one of four institutions and interact only with other group members who adopt the same institution in that period. Migration between institutions is costless, and subjects can adopt any of the institutions at the start of each period. The institutions are (i) No Punishment, (ii) Uncoordinated Peer Punishment, (iii) Coordinated Peer Punishment, and (iv) Coordinated Central Punishment. By "Coordinated," we mean the presence of a normative request that can be used as a coordination device for contributions; it does not refer to the coordination of punishment. These institutions are described in more detail in Section 1.2.1.

Each period contains a contribution stage which is identical to the decision in Part 1, except that contributions to the group project only affect members of the group who adopt the same institution. In addition, a punishment stage is added which provides an additional endowment in each period. In the event that only one subject adopts a particular institution in the period, both the contribution stage and punishment stage endowments go directly to the private account, and the subject is not able to contribute to a group account. This design feature was included because the idea of a public good necessarily involves more than one person benefitting from contributions.

We next describe the four institutions in greater detail. During each period, subjects make several decisions, and their choice opportunities depend partly on the institution adopted. We then describe the information provided to subjects at each decision point. Finally, we describe the material payoffs resulting from subjects' choices.

1.2.1 Institutions

Subjects begin each period by selecting which institution they want to adopt in the current period. For the remainder of the period, subjects only interact with other group members who have also adopted the same institution. The sequence of decisions made during each period is summarized in Figure 2.

No Punishment. No Punishment is identical to Part 1, except that (i) contributions to the group project only affect members of the group who adopt the No Punishment institution and (ii) subjects receive a second endowment in the punishment stage which

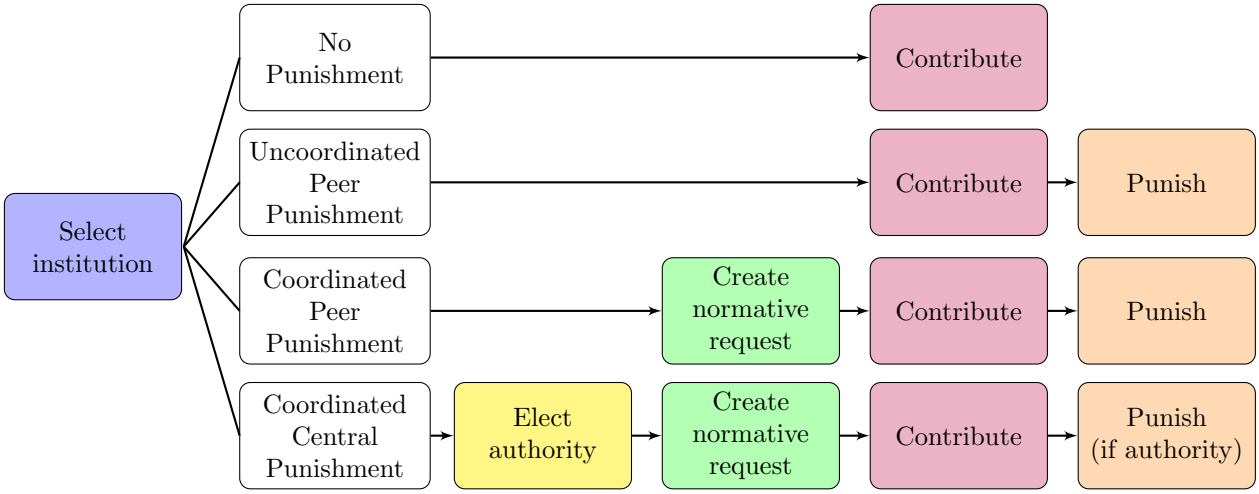


Figure 2: Sequence of decisions in Part 2.

is added directly to their earnings.

Uncoordinated Peer Punishment. In Uncoordinated Peer Punishment, subjects make the same contribution decision as in No Punishment. However, during the punishment stage, subjects can assign deduction points to other institution members which reduce the recipient's earnings at a cost to the person assigning punishment. The cost structure for punishment is described in Section 1.2.3 and defined in equation (4).

Coordinated Peer Punishment. The contribution and punishment stages in Coordinated Peer Punishment are identical to Uncoordinated Peer Punishment. However, prior to the contribution stage, each institution member makes a normative request by privately answering the question, *"How many points do you think each participant should contribute to the project?"*. The average of these private responses is reported on each institution member's computer screen during the contribution stage.[4] The normative request is non-binding and public knowledge within the institution. The cost structure for punishment is described in Section 1.2.3 and defined in equation (4).

Coordinated Central Punishment In Coordinated Central Punishment, one member of the institution is elected to assign all of the punishment for the group, and the total cost of punishment is shared equally by all institution members. At the start of the period, subjects in Coordinated Central Punishment vote for a single institution member

[4] While the normative request is cheap talk, it may act as a coordination device for equilibrium selection when multiple cooperative equilibria exist; social preference models, eg. Fehr and Schmidt (1999), often suffer from the problem of multiple equilibria. In such cases, the normative request is both self-signalling and self-committing in the sense of Farrell and Rabin (1996).

to assign the punishment; the central authority is the person who receives the most votes, and ties are broken randomly. After casting their votes, subjects then create a normative request in the same manner as in Coordinated Peer Punishment. Then, subjects enter the contribution stage, where they are informed of the normative request and make the same contribution decision as in all other institutions. Finally, during the punishment stage, only the central authority can assign deduction points to institution members, and these deduction points reduce the recipient's earnings at a cost which is shared equally by all members of the institution.[5] The cost structure for punishment is described in Section 1.2.3 and defined in equation (5).

Other possible institutions. Our design superficially looks to be a 3×2 factorial design with two omitted institutions, (i) Coordinated No Punishment and (ii) Uncoordinated Central Punishment. We first want to stress that we do not actually have a factorial design, as we do not run separate sessions or treatments for each institution. Instead, we allow subjects to endogenously determine – based on their decisions – whether all institutions, some institutions, or only one institution will be adopted in each period. Our motivation is to understand what institutions people will adopt in natural environments and to see how successful these institutions become in establishing and maintaining cooperation.

In our view, these two potential institutions are not relevant to understanding natural environments. We start from the perspective that peer punishment is always available in natural environments, negating the need to include No Punishment + Norm; in addition, numerical cheap-talk, such as the normative request we use in this paper, typically are not effective in establishing cooperation.[6] The No Punishment institution is a convenient benchmark and has been the traditional way of examining social dilemmas, hence its inclusion. We also do not find it plausible that a group would somehow lose the ability to establish a normative request when moving from peer punishment to centralized

[5]To our knowledge, this particular institution is novel. Therefore, in Online Appendix A.1, we characterize a class of cooperative equilibria in which some individuals have inequity averse social preferences in a one-shot interaction; we do so for comparison with the other institutions based on previous research, eg. Fehr and Schmidt (1999). The existence of only one group member who is strongly averse to inequality is both necessary and sufficient for the existence of cooperative equilibria. Under peer punishment with social preferences, a single group member who is inequity averse is necessary but may not always sufficient for the existence of cooperative equilibria.

[6] Verbal and written communication often enhances cooperation (Isaac and Walker, 1988; Sally, 1995; Ostrom, 1998); however, numerical communication in which written messages cannot be sent is generally unable to establish cooperation and occasionally performs worse than an environment without communication (Wilson and Sell, 1997; Bochet et al., 2006; Bochet and Putterman, 2009). In light of these previous findings, one can reasonably assume that numerical communication itself is not the driving force behind cooperation in our institutions with a cheap-talk norm, though it may interact with and enhance institution performance.

punishment, which negates the need to include Central Punishment without a norm.

We also face a practical concern regarding the number of institutions because we do not allow a subject to contribute to a public good if she is the only one to adopt the institution in the period. If we increase the number of institutions, subjects face a primary concern of adopting an institution based on their beliefs that at least one other person will adopt the same institution; otherwise, they will not be able to even potentially benefit from a public good. Concerns about beliefs in such cases casts doubt on any inferences that can made regarding preferences over institutions, as beliefs about others' choices can override one's own preference. These concerns also led us to omit the two implausible institutions.

1.2.2 Information

The information provided during each period is summarized in Figure 3. All information described in Figure 3 is provided in each institution even if no decision is made at that point. We intentionally chose this feature to rule out desire for increased information as a confounding explanation for institution selection. Subjects receive more information about their own institution than they do about the other institutions, which captures the notion that we know about the people we interact with but only have limited information about those we do not interact with.

At the beginning of each period and before subjects join an institution, all subjects are informed of the number of group members who joined each institution in the previous period and the average earnings for each institution in the previous period. After subjects join an institution, they learn the contribution of each current institution member in the previous period; the Coordinated Central Punishment institution elects the central authority at this point.[7] During the contribution stage, the Coordinated Peer Punishment and Coordinated Central Punishment institutions are informed of the institution's normative request. Finally, at the punishment stage, members of all institutions observe the contribution of each institution member in the current period.

1.2.3 Material payoffs

Earnings in each period are the sum of the contribution stage earnings and the punishment stage earnings. Contribution stage earnings for individual i in institution $inst$, $\pi^1_{i,inst}$, are given by

$$\pi^1_{i,inst} = e^1 - g_{i,inst} + \left(\frac{m}{s_{inst}}\right)\sum_{j=1}^{s_{inst}} g_{j,inst}. \tag{2}$$

[7] In the first period of Part 2 (period 6 overall), subjects are informed of the average contribution of each current institution member during Part 1.

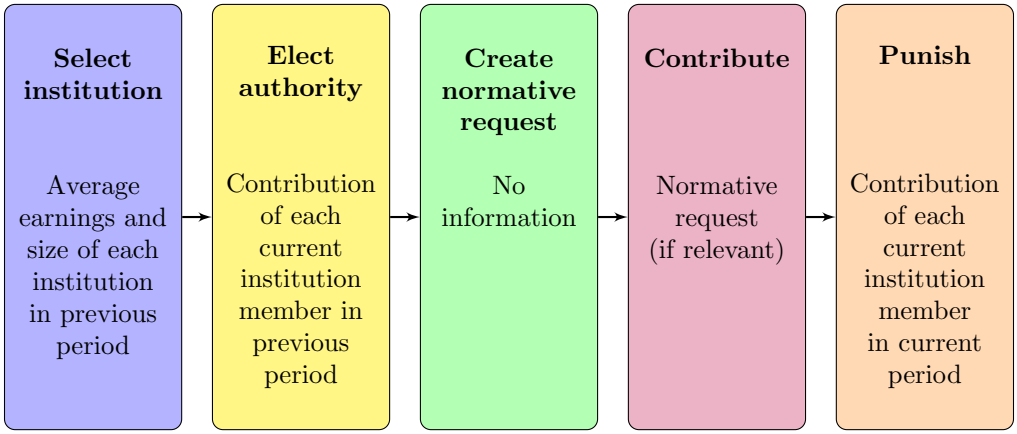


Figure 3: Sequence of information in Part 2. Information is known by all institution members regardless of whether a decision is made at that stage.

where e^1 is the contribution stage endowment, $g_{i,inst}$ is the individual's contribution to the institution project, m is the multiplier for contributions to the institution project, and s_{inst} is the endogenously determined institution size (number of institution members).

Punishment stage earnings for individual i in institution $inst$, $\pi^2_{i,inst}$, are given by

$$\pi^2_{i,inst} = e^2 - c_{i,inst}(d) - r\sum_{j=1}^{s_{inst}} d_{ji,inst} \tag{3}$$

where e^2 is the punishment stage endowment, $c_{i,inst}(d)$ is the cost of assigned punishment, r is the reduction in earnings for each deduction point received, and $d_{ji,inst}$ is the number of deduction points assigned to subject i by subject j in institution $inst$. The cost of assigned punishment, $c_{i,inst}(d)$, differs by institution and is given by

$$c_{i,inst}(d) = \sum_{j=1}^{s_{inst}} d_{ij,inst} \tag{4}$$

if $inst \in$ {Uncoordinated Peer Punishment, Coordinated Peer Punishment} and

$$c_{i,inst}(d) = \frac{\sum_{j=1}^{s_{inst}} d_{Aj,inst}}{s_{inst}} \tag{5}$$

if $inst$ = Coordinated Central Punishment, where $d_{Aj,inst}$ is the number of deduction points assigned to subject j by the central authority A in the Coordinated Central Punishment institution. In the No Punishment institution, $c_{i,inst}$ and $d_{ji,inst}$ are always equal to zero for all i and j.

Finally, we imposed a bankruptcy condition so that earnings in a single period could not be negative. Subjects would still have to pay for assigned deduction tokens even if

Parameter	Value	Meaning
e^1	20	Endowment for contribution stage
m	1.5	Multiplier for contribution to public good
s_{inst}	Endogenous	Institution size
e^2	20	Endowment for punishment stage
r	3	Reduction in earnings from unit of punishment
$c_{i,inst}(d)$	See equations (4) and (5)	Cost of assigned punishment

Table 2: Parameter values used in Part 2.

the tokens could not reduce the recipient's earnings any further. Therefore, per-period earnings are given by

$$\pi_{i,inst} = \max\{\pi^1_{i,inst} + \pi^2_{i,inst}, 0\}. \tag{6}$$

Table 2 summarizes the parameter values used in the experiment. We set $e^1 = e^2 = 20$, $m = 1.5$, and $r = 3$. The total cost of each deduction point is 1, but the cost for each institution member is determined according to equation (4) in institutions utilizing peer punishment and equation (5) in the central punishment institution.

1.2.4 Control treatment – Exogenously assigned perfect matching groups

Under endogenous institution selection, institutions may be successful in establishing and maintaining cooperation for two primary reasons. First, cooperative individuals may select into the same institutions, and cooperation is likely to follow regardless of the institution itself. Second, the institution may create incentives that induce cooperative behavior, regardless of whether the individuals joining the institution are generally cooperative. In our view, the most likely explanation is an interaction of these aspects.Our design with endogenous selection does not allow us to disentangle these explanations. Therefore, we conducted a control treatment with exogenous assignment. For each group in our endogenous selection sessions, we create a matching group of the same size and with identical migration patterns. This exogenous assignment allows us to examine the effects of institutions independently of self-selection.

1.3 Part 3: Social Value Orientation

Part 3 of the experiment consists of the Social Value Orientation (SVO) measure of Murphy et al. (2011) which consists of six allocation decisions between oneself and one other anonymous individual. The allocation decisions constitute modified versions of the

dictator game (Forsythe et al., 1994) in which the relative price of giving varies, similar to the approach in Andreoni and Miller (2002); unlike Andreoni and Miller (2002), the SVO measure provides a numeric score which can be used to make comparisons across individuals. Higher SVO scores indicate greater prosociality. Full details of the SVO measure are provided in Online Appendix A.2. This part of the experiment was conducted online after subjects completed Parts 1 and 2 in the lab. Subjects knew that one of their allocation decisions would be selected and implemented and that they would be the recipient of a different person's allocation decision; payments were mailed to subjects.

1.4 Experimental procedures

Sessions for Part 1 and Part 2 were conducted in a computer lab at the University of Zurich in December 2012 and April 2013. Part 3 was conducted online using Qualtrics (www.qualtrics.com). Experimental instructions are provided in the Online Appendix. Subjects were mostly students from the University of Zurich and Swiss Federal Institute of Technology (ETH-Zurich). Recruitment was conducted using ORSEE (Greiner, 2004), and we excluded students who listed economics or psychology as their major in ORSEE from receiving invitations. Experiments were programmed in z-Tree (Fischbacher, 2007).

Points were used as the experimental currency and converted to Swiss Francs (CHF) at the end of the study; subjects were informed of the exchange rate in the instructions. In Parts 1 and 2, conducted during the same lab session, the exchange rate is 1 point = CHF 0.05. In Part 3, conducted online, the exchange rate is 1 point = CHF 0.10. Average earnings were CHF 56.41 for the lab session (consisting of both Parts 1 and 2), including a show-up fee of CHF 10. Earning from Part 3 were CHF 15-20. Lab sessions lasted 2.5-3 hours on average, and the online portion of the experiment took 10-20 minutes.

Overall, 256 subjects participated in the lab sessions; 128 subjects participated in the endogenous selection treatment, and another 128 subjects participated in the exogenous assignment treatment. Each treatment consisted of eleven groups in total. Nine of these groups had twelve members. One group of nine members and one group of eleven members were used in each treatment due to subjects not coming to the lab. Partner matching was used in Parts 1 and 2, so each group remained fixed for the entire lab session.

For subjects in the endogenous treatments (where we predict assortment effects), 84 of 128 subjects (66%) completed Part 3 of the experiment online. We do not find obvious evidence for selection effects. Subjects who completed Part 3 (average age = 22.0, percent female = 0.37, average earnings in Parts 1 and 2 = CHF 57.40) are similar in observable characteristics to those who did not complete Part 3 (average age = 22.3, percent female = 0.34, average earnings in Parts 1 and 2 = CHF 56.66).

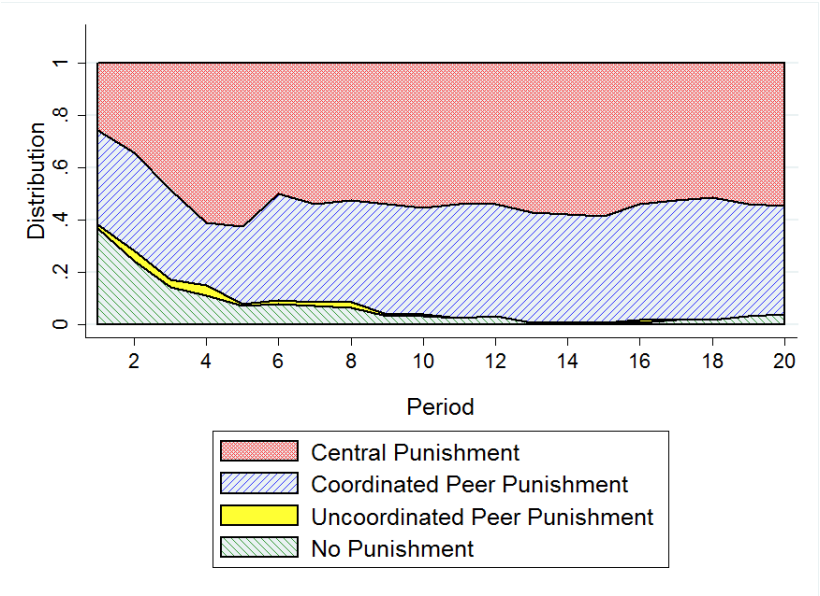


Figure 4: Distribution of subjects across institutions in Part 2.

2 Results

We begin by taking a revealed preference approach to institutions. When subjects are given a choice, which institutions are adopted? The answer provides our first major result and demonstrates that Uncoordinated Peer Punishment is almost never adopted. We include the institution in Result 1 but drop it from subsequent analysis due to lack of observations.

Result 1. *Subjects are initially evenly distributed between the No Punishment (NP), Coordinated Peer Punishment (PP), and Centralized Punishment (CP) institutions while Uncoordinated Peer Punishment is almost never chosen. Over time, only Coordinated Peer Punishment and Centralized Punishment survive.*

Evidence for Result 1. The result is documented in Figure 4, which displays the aggregate distribution of subjects (i.e. data from all groups is pooled to calculate the distribution). The figure shows that initially a substantial number of subjects are in NP, but after period 5, the percentage of subjects entering NP becomes negligible. In addition, the figure shows that, from the very beginning, the share of subjects selecting Uncoordinated Peer Punishment is negligible.

The strong dominance of coordinated peer sanctioning and centralized sanctioning raises the question of why these two institutions prevail. With a revealed preference approach, one would suspect these institutions to prevail because they are successful either in maintaining cooperation or increasing earnings. We find evidence for both conjectures.

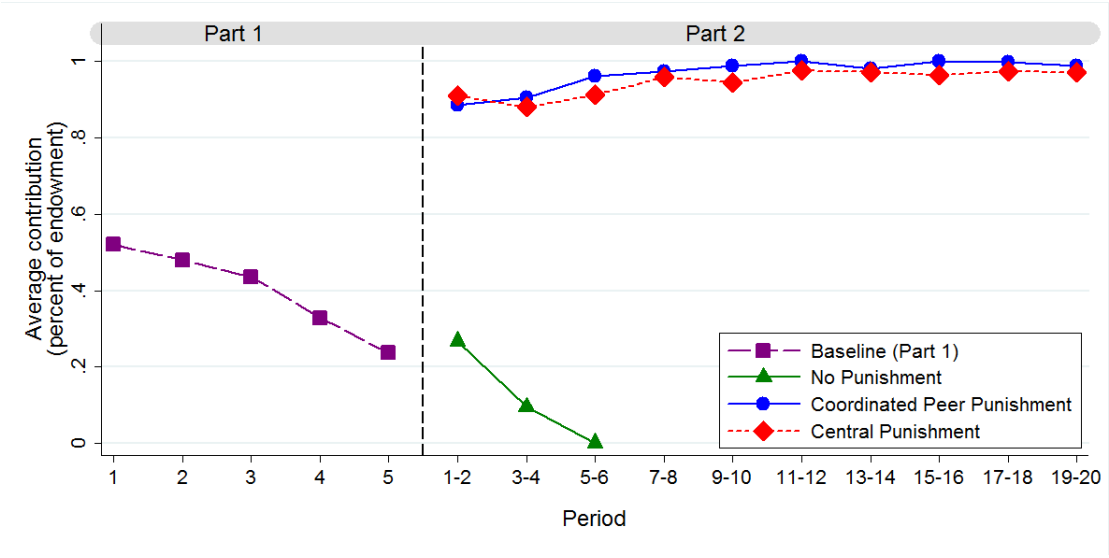
Result 2. *Under both PP and CP, very high cooperation levels are quickly obtained. Initially, only the CP institution outperforms the NP institution in terms of aggregate payoff. However, within a few periods, PP also outperforms the NP institution.*

Evidence for Result 2. The trend for contributions can be seen in Figure 5a. The unit of observation is the matching group, so that the average contribution for an institution is first taken at the group level; the result is then averaged across groups, which is displayed in the figure. Figure 5a shows that already in the first period of Part 2 the average contribution (as a percentage of the endowment) in PP and CP is about 90% and soon reaches close to 100%. In contrast, average contributions in NP quickly decline to low levels. Evidence for earnings is provided in terms of efficiency, which we define as the percentage of the social optimum (full contributions by all subjects) earned by subjects in the institution. Average efficiency can be seen in Figure 5b. In the first period of Part 2, efficiency is significantly higher in CP relative to NP (Mann-Whitney U, p=0.001). However, efficiency in PP is not significantly different than in NP during the initial period (Mann-Whitney U, p=0.066). By period 4, the efficiency in PP is significantly higher than in the NP.

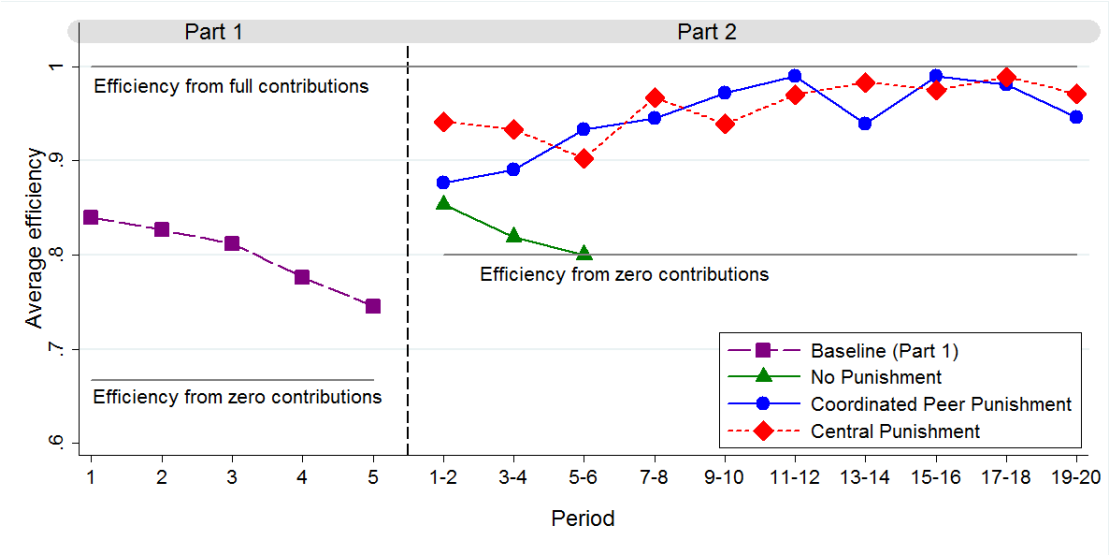
Why do coordinated peer punishment and centralized punishment perform so well in terms of quickly establishing and maintaining cooperation and in terms of aggregate payoffs? One reason could be that the institutions themselves have beneficial intrinsic properties that lead to high performance – at least in the long run – regardless of the migration pattern (i.e. even with random assignment to institutions). A second reason could be that prosocial individuals are the first ones to migrate into these institutions and quickly establish a beneficial social norm of full cooperation such that whose who join later can be smoothly integrated into a "high cooperation society." In the following, we study these two potential factors.

Result 3. *Under exogenous assignment, institutions that allow for sanctioning and normative requests induce cooperative behavior. These institutions perform better in the long run than NP under exogenous assignment. Under both endogenous selection and exogenous assignment, the normative request is an effective coordination device for contributions, and contributions increase immediately upon entering the sanctioning regimes from the non-sanctioning institution.*

For the sake of clarity, we examine the evidence for each of these properties separately. We first examine the effectiveness of PP and CP under exogenous assignment. Selection effects cannot explain superior performance of these institutions because subjects are exogenously assigned to an institution.



(a) Average contribution.

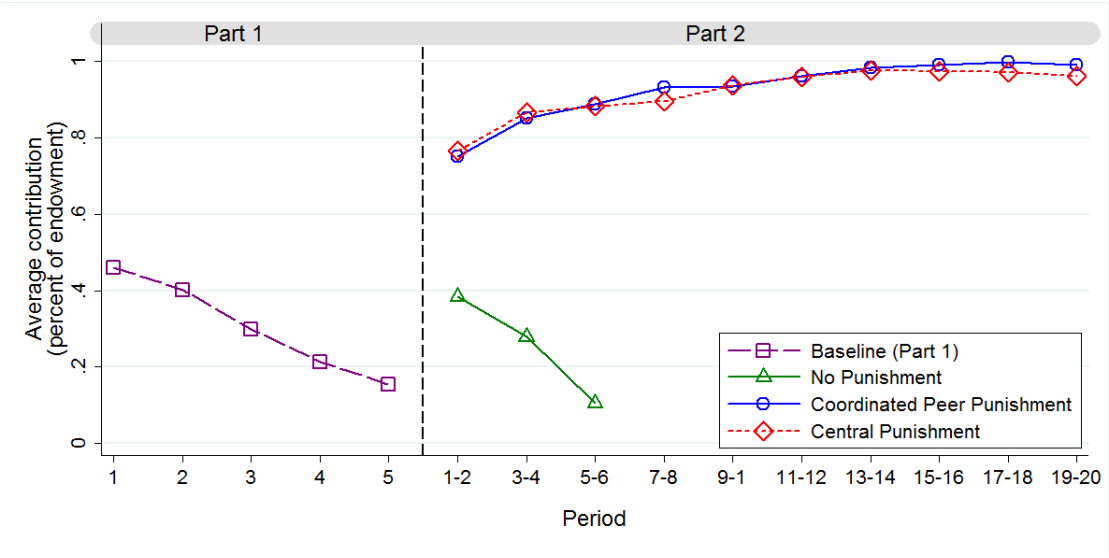


(b) Average efficiency.

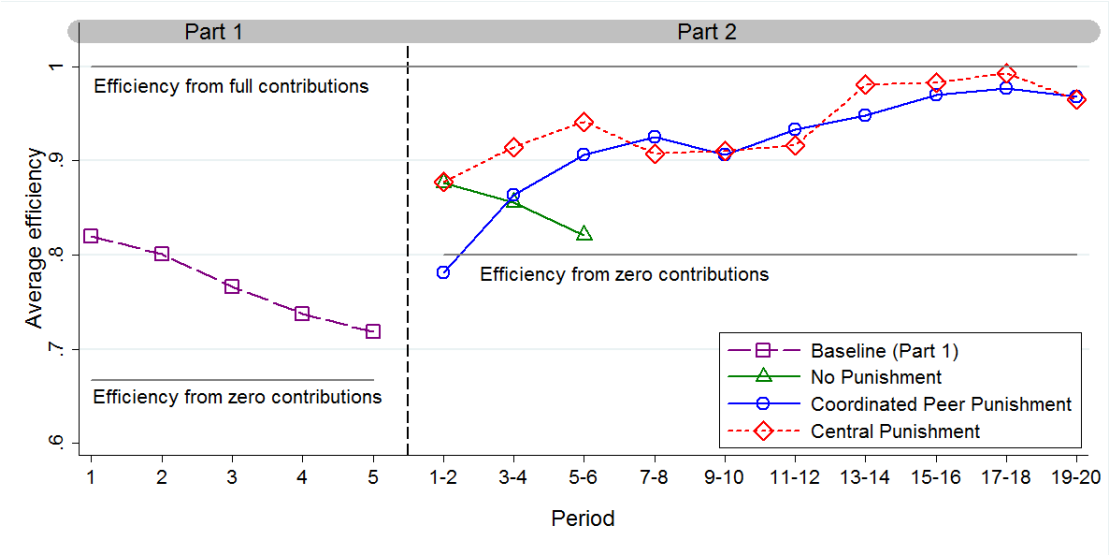
Figure 5: Contributions and efficiency under endogenous selection. The unit of observation is the matching group, so that the average for an institution is first taken at the group level; the result is then averaged across groups, which is displayed in the figure. Efficiency is the percentage of the socially optimal outcome (full contributions) earned by subjects. Efficiency from zero contributions differs between periods 1–5 (Part 1) and periods 6-25 (Part 2) due to the additional endowment received in periods 6–25.

Result 3a. *Institutions that allow for sanctioning and normative requests perform better than NP after a few periods. Initially, the efficiency in PP and CP is not higher than in NP; however, after a few periods, both PP and CP outperform NP.*

Evidence for Result 3a. The trend for contributions can be seen in Figure 6a. Figure 6a shows that average contributions as a percentage of the endowment are roughly 75%



(a) Average contribution.



(b) Average efficiency.

Figure 6: Contributions and efficiency under exogenous assignment. The unit of observation is the matching group, so that the average for an institution is first taken at the group level; the result is then averaged across groups, which is displayed in the figure. Efficiency is the percentage of the socially optimal outcome (full contributions) earned by subjects. Efficiency from zero contributions differs between periods 1–5 (Part 1) and periods 6-25 (Part 2) due to the additional endowment received in periods 6–25.

already in the first period of Part 2, and they reach 100% after 12 periods. Thus, contributions are significantly higher in the punishment institutions than in NP from the beginning for both PP (Mann-Whitney U, p=0.066) and CP (Mann-Whitney U, p=0.001). Average efficiency is documented in Figure 6b, which shows that, during the first few periods of Part 2, the efficiency in the NP treatment is rather similar compared

to the PP and the CP; however, from periods 5-6 onwards, efficiency in both PP and CP is significantly higher than in NP. These results indicate that under exogenous assignment there are considerable initial efficiency losses due the punishment activities of the players.

The aggregate pattern of contributions and efficiency in our coordinated peer punishment institution under exogenous assignment is remarkably similar to previous studies on uncoordinated peer punishment, in which the institution induces high contributions but initially performs worse than a sanction-free environment due to the use of punishment (Gächter et al., 2008). The existence of a strong institutional effect on contribution levels suggests that we should also observe changes in contribution behavior when individuals migrate into PP and CP from NP. Indeed, we do observe these changes.

Result 3b. *Individual contributions never decrease when individuals migrate from NP to either PP or CP, and most people increase their contributions upon entering a sanctioning regime.*

Evidence for Result 3b. Figure 7 displays subjects' contribution types upon entering a punishment institution as a function of contribution type in the period immediately before entering the punishment institution. For example, approximately 70% of the subjects who were low contributors in endogenous selection (far left in Figure 7) become high contributors upon entering a punishment institution. Taken together, Figure 7 documents that even many of those subjects who were low contributors before they entered a punishment institution immediately turned into higher contributors upon entering a punishment institution.

We acknowledge that while this behavioral change provides evidence for institutional effects, the result itself is not terribly surprising. It is well documented in the literature that institutions allowing for punishment induce changes in behavior, even within the same subject (Fehr and Gächter, 2000; Gürerk et al., 2011). The noteworthy aspect of the institutions adopted by subjects is the use of the normative request. While the threat of punishment may increase contributions, there is still now way of knowing the "right" contribution and what others expect you to give. The normative request, despite being cheap talk, solves this problem, even though the request itself is not generally effective in the absence of punishment.[8]

Result 3c. *The normative request is an effective coordination device for contributions in PP and CP, and it is effective under both endogenous selection and exogenous assignment.*

[8]See footnote 6 for discussion and references on numerical communication in public goods experiments.

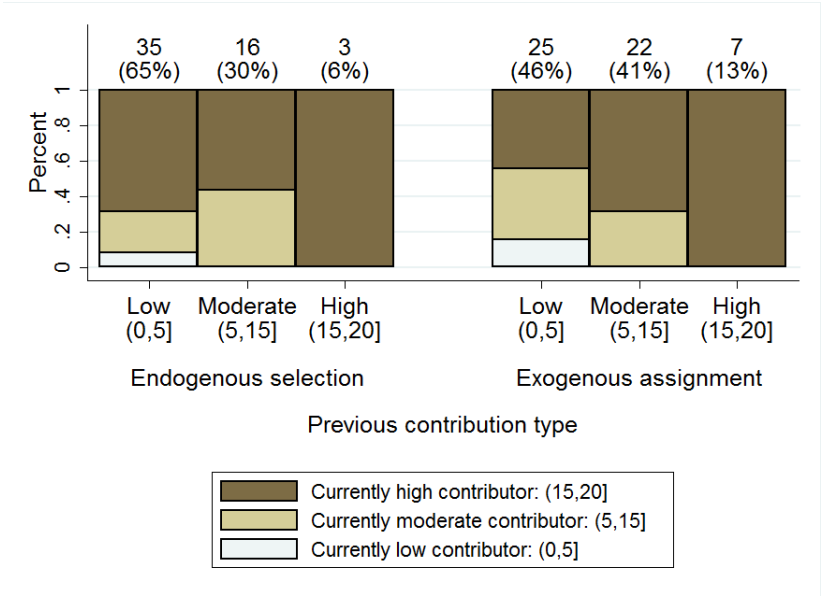


Figure 7: Contribution behavior immediately before and after entering a punishment institution. Horizontal axis indicates the contribution type in the No Punishment institution in the last period before joining a punishment institution. Bar height indicates relative frequency of contribution type upon entering a punishment institution conditional on type before migrating. Number above bars indicates the overall number of subjects from that treatment who fall into the category.

Evidence for Result 3c. We restrict attention to the first five periods of Part 2 because average contributions tend towards full contributions quickly and eliminate any variation in the data afterwards. To provide support for Result 3c, we regress subjects' contributions levels on the average normative request in a group and a constant and a restricted model in which we set the coefficient on the constant term to zero (i.e. $\beta_0 = 0$); these models are reported in Table 3.[9] We can only reject the restricted model in favor of

[9]An econometric issue arises when running regressions on our data, but there is a simple solution. We need to cluster standard errors at the level of the matching group, and we have a relatively small number of matching groups. It is well known that in such cases the standard errors will be inconsistent and lead to over-rejection of the null hypothesis (Bertrand et al., 2004). Bootstrapping can overcome this problem, and we use a pairs cluster bootstrap-t with cluster robust standard errors (Cameron et al., 2008) that performs quite well in their simulations. In this bootstrap, resampling is at the level of clusters instead of individual observations. The Wald statistics from the bootstrap samples are then used to create the distribution against which the Wald statistic from the original data is tested. Cameron et al. (2008) suggests using a wild cluster bootstrap-t procedure in OLS regressions; however, in their simulations, the pairs cluster bootstrap-t with cluster robust standard errors performs only slightly worse than the wild cluster bootstrap-t. We opt to use the pairs cluster bootstrap-t with cluster robust standard errors for consistency with later analyses in which the wild cluster bootstrap-t is inappropriate, specifically the Tobit regressions in Table 4. The wild bootstrap involves estimating $\hat{y}_i$ and either adding or subtracting the residual ε_i with equal probability to create new pseudo-samples, where $\hat{y}_i^*$ is the value of y_i in the wild bootstrap pseudo-sample. With corner solutions in a Tobit regression, we would need to set $\hat{y}_i^* = 0$ whenever $\hat{y}_i^* < 0$. In doing so, we would have $\hat{\beta}_i^* \neq \hat{\beta}_i$, which appears inappropriate for inference using the wild bootstrap.

	Endogenous Selection		Exogenous Assignment	
	Peer Punishment	Central Punishment	Peer Punishment	Central Punishment
Panel A – Restricted model (constant constrained to $\beta_0 = 0$)				
Normative request	0.981 (0.000)***	0.997 (0.000)***	1.001 (0.000)***	0.968 (0.000)***
Panel B – Unrestricted model				
Normative request	1.078 (0.000)***	0.686 (0.000)***	1.144 (0.000)***	1.073 (0.000)***
Constant	-1.827 (0.083)*	5.693 (0.581)	-2.253 (0.001)***	-1.803 (0.000)***
Likelihood ratio, $\chi^2(1)$	0.54 (0.461)	6.69 (0.010)***	2.49 (0.115)	1.69 (0.194)
N	201	292	201	292
Number of clusters	11	11	11	11
Bootstrap samples	9,999	9,999	9,999	9,999

Table 3: The role of normative requests as a coordination device for contributions in the first five periods (periods 6-10 overall). OLS regressions with robust standard errors clustered by matching group. Bootstrapped p-values given in parentheses based on 9,999 bootstrap samples computed using pairs cluster bootstrap-t with clustered standard errors.

a model with both normative request and constant for CP under endogenous selection. However, for commonly observed values of the normative request (16-20), the predicted contributions almost perfectly match the normative request.[10] For the remaining conditions, we cannot reject the model with only the normative request as a regressor using a likelihood ratio test. In the restricted model, coefficients on the normative request range from 0.968 to 1.001, indicating that normative request is a clear coordination device leading to contributions that are not significantly different from the normative request.

We have so far provided evidence that coordinated peer punishment and centralized punishment institutions have beneficial intrinsic properties that lead to high performance – at least in the long run – regardless of the migration pattern (i.e. even with random assignment to institutions). However, these intrinsic properties were one of the two potential reasons for why these institutions perform so well in terms of quickly establishing

[10]For example, with a normative request of 16, the predicted contribution is 16.669 ($= 5.693 + 16 \times 0.686$). With a normative request of 20, the predicted contribution is 19.413 ($= 5.693 + 20 \times 0.686$).

and maintaining cooperation and in terms of aggregate payoffs. A second reason could be that prosocial individuals are the first ones to migrate into these institutions and quickly establish a beneficial social norm of full cooperation such that those who join later can be smoothly integrated into a "high cooperation society." In the next two results, we study the latter potential factor and find evidence in support of sorting effects.

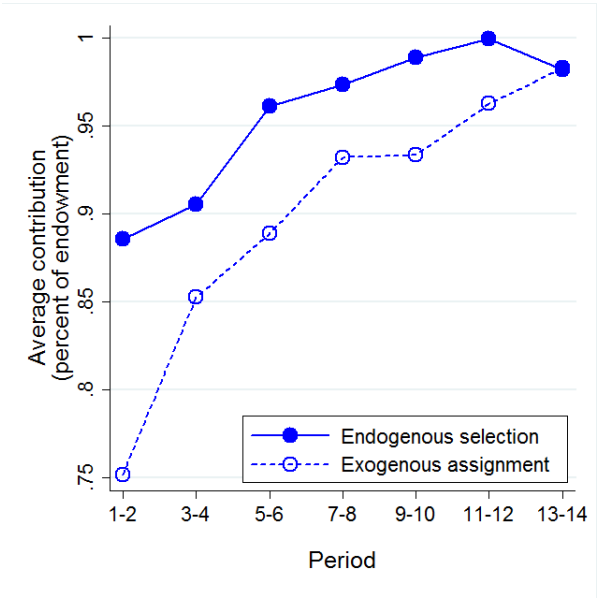
Result 4. *Under endogenous selection, prosocial individuals are quick to migrate into the coordinated peer punishment and centralized punishment institutions and establish a culture of high cooperation. This culture of cooperation includes making higher normative requests and following through by making higher contributions than the subjects under exogenous assignment.*

Evidence for Result 4. We provide four pieces of evidence in support of the result. Statistical results presented as evidence are based on Mann-Whitney U tests.

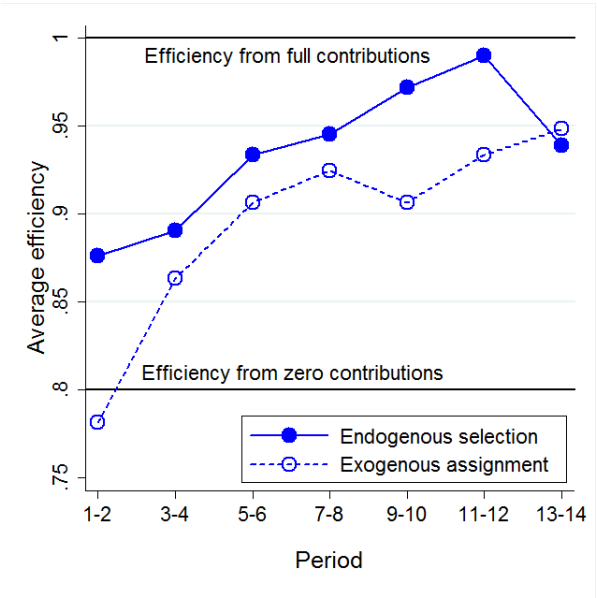
1. Contributions and efficiency are much higher in PP and CP under endogenous selection relative to exogenous assignment. While the institutions under exogenous assignment eventually converge to the performance under endogenous selection, it takes much longer. The trend across time can be seen in Figure 8. In the first period of Part 2, the endogenously selected institutions perform significantly better with respect to both contributions (PP, p=0.021; CP, p=0.012) and efficiency (PP, p=0.016; CP, p=0.035). The unit of observation is the matching group, so that one observation is obtained for each institution in a matching group (if it is adopted).

2. Higher average contributors during the baseline public goods game in Part 1 of the experiment are more likely to adopt coordinated peer punishment and centralized punishment institutions in the first period of endogenous selection. This behavior can be seen clearly in Figure 9a, where the average individual contribution during the baseline public goods game without punishment in Part 1 is the unit of analysis (PP > NP, p=0.012; CP > NP, p=0.015).

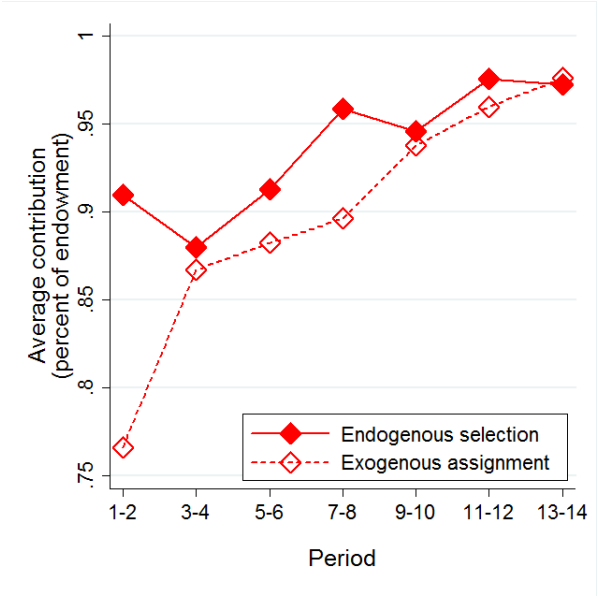
3. Prosocial individuals, as measured by the Social Value Orientation scale in Part 3 of the experiment (conducted online), are more likely to adopt coordinated peer punishment and centralized punishment institutions in the first period of endogenous selection. This behavior can be seen in Figure 9b. Individual SVO scores are the units of analysis. Since not all subjects completed Part 3 of the experiment online, we pool the PP and CP institutions for the statistical test and find that more the prosociality of individuals selecting into punishment institutions is significantly



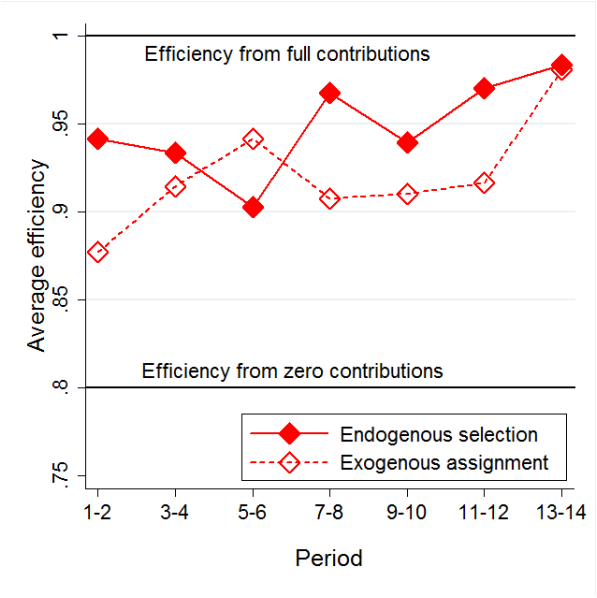
(a) Contributions in coordinated peer punishment.



(b) Efficiency in coordinated peer punishment.

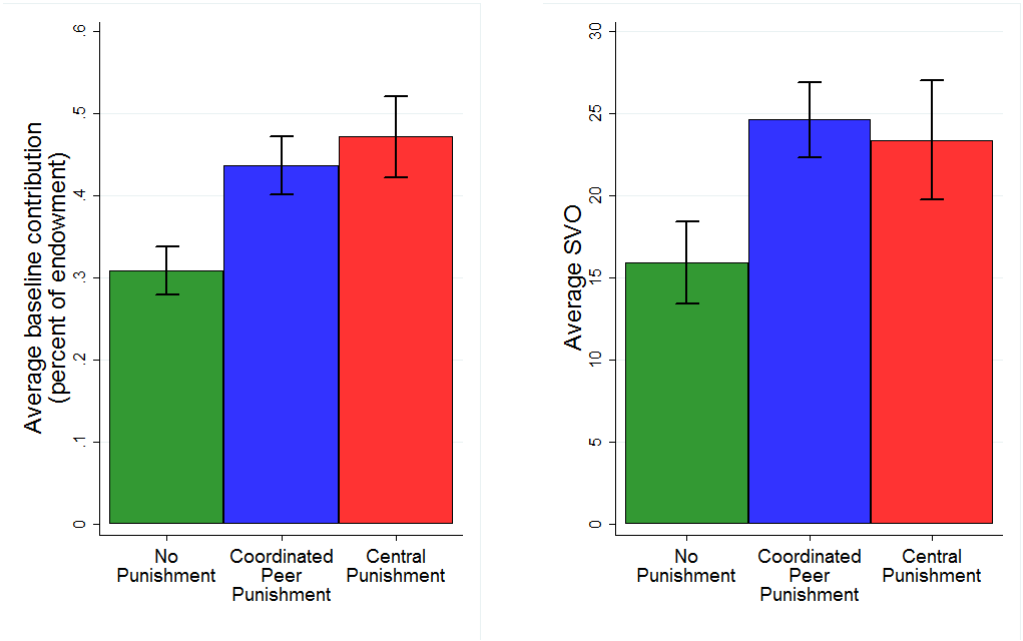


(c) Contributions in central punishment.



(d) Efficiency in central punishment.

Figure 8: Contributions and efficiency between treatments in Part 2. The unit of observation is the matching group, so that the average for an institution is first taken at the group level; the result is then averaged across groups, which is displayed in the figure. Efficiency is the percentage of the socially optimal outcome (full contributions) earned by subjects. Periods 15-20 are omitted from the figure since there is no observable difference between treatments for either institution in these periods.



(a) Average baseline contribution

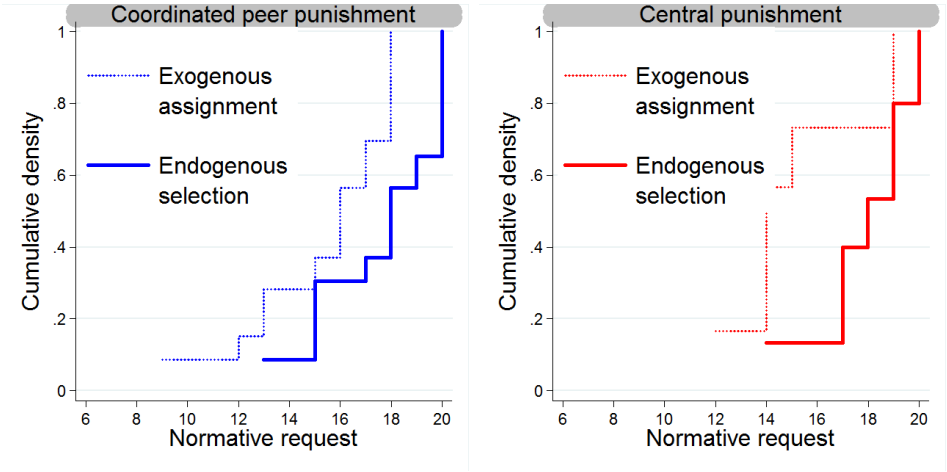
(b) Social Value Orientation

Figure 9: Assortment in first period of endogenous selection (period 6 overall). Error bars represent standard error of the mean.

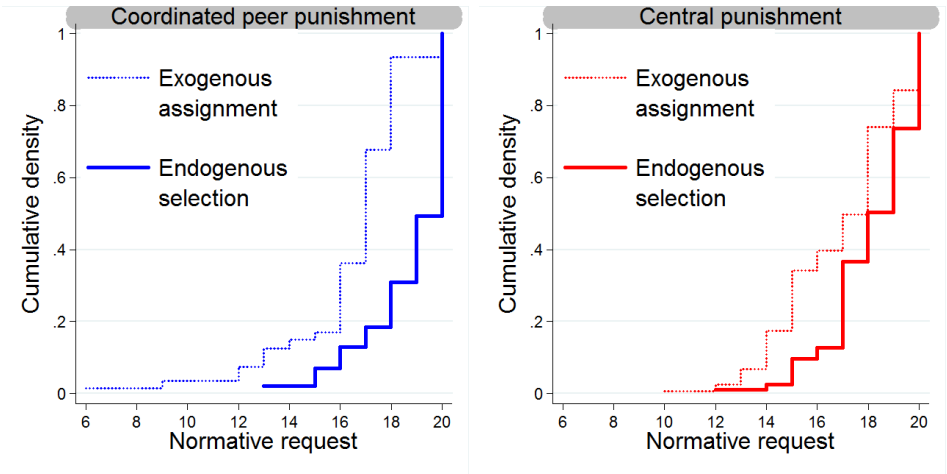
higher than those who select into NP (PP and CP > NP, p=0.021).[11]

4. Individuals in the coordinated peer punishment and centralized punishment institutions make higher normative requests than members of the same institution under exogenous assignment in the early periods. Figure 10 displays the cumulative density functions for individual normative requests in both institutions under endogenous selection and exogenous assignment for (a) the first period and (b) the first five periods. In all cases, the normative request under endogenous selection first-order stochastically dominates the normative request under exogenous assignment. In the first five periods of the endogenous selection treatment more than 80% of the subjects in PP had a normative request of 18 or more while only about 30% of the subjects had similarly high requests under exogenous assignment. Likewise, almost 90% of subjects in the CP institution requested a contribution level of 17 or more when they self-selected into this institution while only 60% had similarly high requests under exogenous assignment. These results show that the vast majority of subjects requested very high contribution levels right from the beginning under

[11]See Section 1.4 for lack of evidence for bias in completing Part 3. For those who completed Part 3, 30 subjects selected NP in the first period of Part 2, 33 selected PP, and 19 selected CP. Given the similar baseline contributions (Figure 9a) and contributions in the initial period of Part 2 (Figure 5a), we find it reasonable to pool the SVO scores across the two punishment institutions for the statistical test.



(a) Period 1.



(b) Periods 1-5.

Figure 10: Cumulative density functions of normative requests in (a) the first period and (b) the first five periods. In all cases, the normative request under endogenous selection first-order stochastically dominates the normative request under exogenous assignment. Density functions are based on individual normative requests in each period.

endogenous selection, but a substantial number of subjects were satisfied with lower requests under exogenous assignment.

Taken together, these results suggest that subjects adopting punishment institutions under endogenous selection quickly established a cooperative culture. These institutions attracted a high share of prosocial individuals who quickly established high normative requests; these requests successfully coordinated the whole group to high contribution levels such that little punishment was necessary to enforce the widely agreed high contribution norm. In contrast, exogenous assignment of subjects to institutions puts sand into the gears of cooperation. It prevents the self-selection of prosocial individuals and

causes substantial adjustment costs during the initial phases because subjects demand lower contributions and cooperate less, which then requires higher punishment costs to establish cooperation.

Why can't the subjects under exogenous assignment coordinate on high normative requests in the early periods? One reason could be that subjects in the exogenous assignment treatments are aware that the institution members are not very cooperative; therefore, trust must be built up over time and incrementally from lower initial requests. While this question is interesting, we cannot address it with our data; thus, it remains open for future investigation.

One final result emerged from our data that we did not anticipate in advance. Our centralized punishment institution was motivated by the presence of such institutions across societies from small-scale tribes to international organizations such as the United Nations Security Council, and we anticipated that centralization would lower the overall amount of punishment by removing the problem of coordinating punishment without communication under peer punishment. Antisocial punishment, in which above-average contributors are punished, is commonly observed under peer punishment (Herrmann et al., 2008). We find that centralization has an important effect on antisocial punishment.

Result 5. *Centralization eliminates antisocial punishment.*

Evidence for Result 5. Table 4 contains average partial effects from Tobit regressions estimating the effects of both negative and positive deviations from the average contribution. In all cases, larger negative deviations from the average are met with significantly higher amounts of punishment. Antisocial punishment remains a problem in coordinated peer punishment, as larger *positive* deviations are met with significantly higher amounts of punishment (average partial effect is 0.212 under endogenous selection and 0.146 under exogenous assignment). However, the opposite is found in centralized punishment, where larger positive deviations are met with significantly lower amounts of punishment (average partial effect is -0.097 under endogenous selection and -0.410 under exogenous assignment). We also considered the possibility that deviations from the normative request were the basis for punishment, but this model performed substantial worse than deviations from the average contribution. We report differences in Akaike Information Criterion, ΔAIC, in Table 4; these values ranged from 26.04 to 79.38 in our data, and $\Delta AIC > 10$ is generally considered very strong evidence in support of the favored model (Burnham and Anderson, 2002).

Why should centralization eliminate antisocial punishment? Some recent papers do not find the same result (Fischer et al., 2013; Grechenig et al., 2013). The critical difference appears to be that we allow institution members to select the central authority,

	Endogenous Selection		Exogenous Assignment	
	Peer Punishment	Central Punishment	Peer Punishment	Central Punishment
Own negative deviation from average contribution	0.234 (0.000)***	0.080 (0.001)***	0.367 (0.000)***	0.195 (0.001)***
Own positive deviation from average contribution	0.212 (0.000)***	-0.097 (0.022)**	0.146 (0.029)**	-0.410 (0.003)***
N	1,008	1,323	1,008	1,323
Left-censored	886	1,194	825	1,167
Uncensored	122	129	183	156
Number of clusters	11	11	11	11
Log-likelihood	-582.82	-589.61	-778.60	-821.82
AIC	1171.64	1185.22	1563.20	1649.64
ΔAIC	79.38	26.04	96.24	41.22
Bootstrap samples	9,999	9,999	9,999	9,999

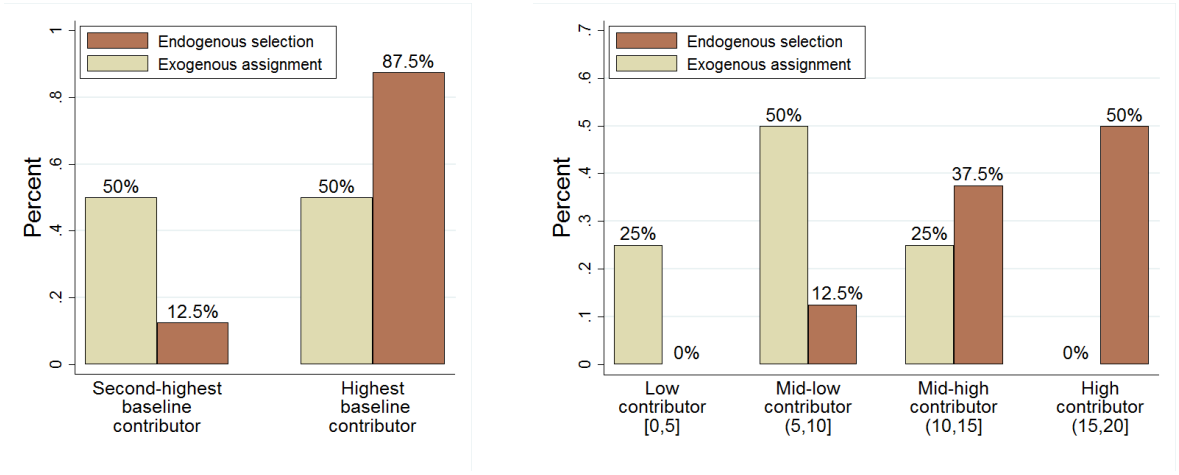
Table 4: Punishment received based on deviation from average group contribution during the period. Tobit regressions with robust standard errors clustered by matching group. Coefficients reported are average partial effects. Bootstrapped p-values given in parentheses based on 9,999 bootstrap samples computed using pairs cluster bootstrap-t with clustered standard errors. ΔAIC is the difference in Akaike Information Criterion for the reported regression compared to a model with deviations from normative request as independent variables; values exceeding 10 are considered very strong evidence in support of the regressions presented in the table.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

while these other recent papers exogenously assign the role randomly and keep it fixed throughout the experiment. In particular, it is often speculated that antisocial punishment comes from below-average contributors targeting above-average contributors either to send a message not to punish low contributions or to retaliate against received punishment (Herrmann et al., 2008). In our experiments, institution members tended to delegate authority to high average contributors, suggesting that they chose the "right" people and did not elect antisocial punishers. Figure 11 illustrates delegation to high contributors in the first period of Part 2. Authority was always delegated to one of the two highest baseline contributors (Figure 11a); the highest contributor was elected 87.5% of the time under endogenous selection but only 50% of the time under exogenous assignment. Figure 11b illustrates that these contribution types, however, are very different between treatments. Under endogenous selection, 50% of the initial authorities contributed at least 16 points to the group project on average during Part 1 and 87.5%



(a) Contribution rank

(b) Contribution type

Figure 11: Baseline contribution behavior (Part 1) for subjects elected as central authority in first period of Part 2 (period 6 overall).

contributed at least 11 points on average. In contrast, 75% of the initial authorities in the exogenous assignment treatment contributed half of their endowment (10 points) or less on average. Thus, while subjects tend to delegate authority to the highest contributors in a relative sense (Figure 11a), there is a difference in absolute contributions across treatments which is due to the self-selection of prosocial individuals in the endogenous selection treatment.

3 Discussion and concluding remarks

Institutions "are the humanly devised constraints that shape human interaction" (North, 1990). People routinely make decisions that determine who they interact with and in which institutions those interactions occur. As a result of these endogenous selection effects and varying institutions, real-world interactions may differ substantially from behavior in typical lab experiments in which interaction partners are random, institutions are exogenously imposed, or – most commonly – both.

There has been a recent surge in the economics literature exploring the role of endogenous group formation on cooperation and a corresponding surge exploring the effect of endogenous institution formation within fixed groups. By only varying one of these two aspects, the other factor, which commonly occurs outside the lab and which may be critical to sustain cooperation, is overlooked; while this approach is useful for identification of a single effect, it also prevents a fuller understanding of cooperation outside the lab. Moreover, the institutional options are often very limited and may not reflect

institutions that are adopted in natural environments.

Our goal in this paper is to provide subjects with the opportunity to select into institutional environments in a public goods game and only interact with others who also select into the same institution. Selection occurs every period, allowing us to observe migration patterns and changes in behavior across time. We allow subjects to select into four institutions. The first two institutions, no punishment and costly peer punishment, are commonly used in lab experiments. In the third institution, we add a cheap-talk norm normative request to peer punishment. In the fourth, we maintain the cheap-talk normative request but now allow the institution members to elect one member to assign all punishment that period but socialize the cost so that the burden is shared equally by all institution members.

Our first major result provides new evidence on which institutions people will adopt when given the option. Subjects are initially evenly divided among the institutions with no punishment, peer punishment with norm, and central punishment with norm; the peer punishment institution used in many studies is almost never adopted. This result questions the inferences we can make about behavior outside the lab based on most previous experiments, as these experiments have typically relied on institutions that subjects would not normally adopt.

We also find several other interesting results. We demonstrate that prosocial individuals are more likely to select into punishment institutions in early periods. We also find that the cheap-talk normative request in these punishment institutions is used as a coordination device for contributions but not for determining punishment, which appears to be driven by deviations from the average contribution. The central punishment institution is able to eliminate the initial inefficiency that is usually observed with peer punishment and is also able to eliminate antisocial punishment.

Our work demonstrates the ability to capture several complex phenomena related to cooperation in an experimental design that allows these phenomena to be disentangled in the lab. The results shed new light on the early stages of institutional emergence and cooperation and how they coevolve in a manner that describes many real-world settings. Our design leaves open the possibility of many extensions that incorporate realistic extensions of the coevolutionary process, of which we suggest a few obvious candidates: costly migration and entry restrictions; permanent formal authorities who can extract rents; increasing group size and imperfect information; and the process of internalizing norms as the basis for punishment. These extensions provide fertile ground for new experiments and allow for a better understanding of the coevolution of cooperative behavior and institutions in natural environments.

References

Acemoglu, D., D. Cantoni, S. Johnson, and J. A. Robinson (2011): "The Consequences of Radical Reform: The French Revolution," *American Economic Review*, 101, 3286–3307.

Ahn, T. K., R. M. Isaac, and T. C. Salmon (2008): "Endogenous group formation," *Journal of Public Economic Theory*, 10, 171–194.

——— (2009): "Coming and going: Experiments on endogenous group sizes for excludable public goods," *Journal of Public Economics*, 93, 336–351.

Andreoni, J. and J. Miller (2002): "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70, 737–753.

Berkowitz, D., K. Pistor, and J.-F. Richard (2003a): "Economic development, legality, and the transplant effect," *European Economic Review*, 47, 165–195.

——— (2003b): "The transplant effect," *American Journal of Comparative Law*, 51, 163–203.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics*, 119, 249–275.

Bochet, O., T. Page, and L. Putterman (2006): "Communication and punishment in voluntary contribution experiments," *Journal of Economic Behavior and Organization*, 60, 11–26.

Bochet, O. and L. Putterman (2009): "Not just babble: Opening the black box of communication in a voluntary contribution experiment," *European Economic Review*, 53, 309–326.

Boehm, C., C. Antweiler, I. Eibl-Eibesfeldt, S. Kent, B. M. Knauft, S. Mithen, P. J. Richerson, and D. S. Wilson (1996): "Emergency Decisions, Cultural-Selection Mechanics, and Group Selection [and Comments and Reply]," *Current Anthropology*, 37, 763–793.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 166–193.

Burnham, K. P. and D. R. Anderson (2002): *Model selection and multi-model inference: a practical information-theoretic approach*, Springer.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414–427.

Charness, G. and M. Rabin (2002): "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, 117, 817–869.

Dreber, A., D. G. Rand, D. Fudenberg, and M. A. Nowak (2008): "Winners don't punish," *Nature*, 452, 348–351.

Farrell, J. and M. Rabin (1996): "Cheap talk," *Journal of Economic Perspectives*, 10, 103–118.

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

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114, 817–868.

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

Fischbacher, U., S. Gächter, and E. Fehr (2001): "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, 71, 397–404.

Fischer, S., K. R. Grechenig, and N. Meier (2013): "Cooperation under Punishment: Imperfect Information Destroys it and Centralizing Punishment Does Not Help," *MPI Collective Goods Preprint*.

Forsythe, R., J. L. Horowitz, N. E. Savin, and M. Sefton (1994): "Fairness in simple bargaining experiments," *Games and Economic behavior*, 6, 347–369.

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

Grechenig, K., A. Nicklisch, and C. Thöni (2013): "Information-sensitive Leviathans – the emergence of centralized punishment," Unpublished working paper.

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments," in *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung*, ed. by K. Kremer and V. Macho, 79–93.

Gürerk, Ö., B. Irlenbusch, and B. Rockenbach (2006): "The competitive advantage of sanctioning institutions," *Science*, 312, 108–111.

——— (2011): "Voting with feet: community choice in social dilemmas," Unpublished working paper.

——— (2013): "On cooperation in open communities," Unpublished working paper.

Hamman, J. R., R. A. Weber, and J. Woon (2011): "An experimental investigation of electoral delegation and the provision of public goods," *American Journal of Political Science*, 55, 738–752.

Henrich, J., J. Ensminger, R. McElreath, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, et al. (2010): "Markets, religion, community size, and the evolution of fairness and punishment," *Science*, 327, 1480–1484.

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, et al. (2006): "Costly punishment across human societies," *Science*, 312, 1767–1770.

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

Isaac, R. M. and J. M. Walker (1988): "Communication and free-riding behavior: The voluntary contribution mechanism," *Economic Inquiry*, 26, 585–608.

Kaplan, H., M. Gurven, K. Hill, and A. M. Hurtado (2005): "The natural history of human food sharing and cooperation: a review and a new multi-individual approach to the negotiation of norms," in *Moral sentiments and material interests: The foundations of cooperation in economic life*, ed. by H. Gintis, S. Bowles, R. Boyd, and E. Fehr, 75–113.

Kimbrough, E. O., V. L. Smith, and B. J. Wilson (2008): "Historical Property Rights, Sociality, and the Emergence of Impersonal Exchange in Long-Distance Trade," *American Economic Review*, 98, 1009–1039.

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

Marlowe, F. W., J. C. Berbesque, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, J. Ensminger, M. Gurven, E. Gwako, J. Henrich, et al. (2008):

"More 'altruistic' punishment in larger societies," *Proceedings of the Royal Society B: Biological Sciences*, 275, 587–592.

Mathew, S. and R. Boyd (2011): "Punishment sustains large-scale cooperation in prestate warfare," *Proceedings of the National Academy of Sciences*, 108, 11375–11380.

Murphy, R. O., K. A. Ackermann, and M. J. Handgraaf (2011): "Measuring Social Value Orientation," *Judgment and Decision Making*, 6, 771–781.

North, D. C. (1990): *Institutions, institutional change and economic performance*, Cambridge university press.

Nunn, N. (forthcoming): "Historical development," in *Handbook of Economic Growth, Volume 2*, ed. by P. Aghion and S. N. Durlauf, North-Holland.

Ostrom, E. (1998): "A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997," *American Political Science Review*, 1–22.

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

Sally, D. (1995): "Conversation and Cooperation in Social Dilemmas A Meta-Analysis of Experiments from 1958 to 1992," *Rationality and society*, 7, 58–92.

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

Tiebout, C. M. (1956): "A pure theory of local expenditures," *Journal of Political Economy*, 64, 416–424.

Wiessner, P. (2005): "Norm enforcement among the Ju/'hoansi Bushmen," *Human Nature*, 16, 115–145.

Wilson, R. K. and J. Sell (1997): ""Liar, Liar..." Cheap Talk and Reputation in Repeated Public Goods Settings," *Journal of Conflict Resolution*, 41, 695–717.

A Online Appendix – Not for publication

A.1 Cooperative equilibria in Central Punishment + Norm

Here, we characterize one set of cooperative equilibria in a one-shot version of the public goods game with centralized punishment and cheap-talk norm. We use the inequity aversion model of Fehr and Schmidt (1999) for simplicity and tractability. One could alternatively use other popular inequity aversion models such as Bolton and Ockenfels (2000) or Charness and Rabin (2002). With Charness and Rabin (2002) preferences, one needs to include their *demerit profile* to capture reciprocity; the simpler version with only the disinterested social-welfare criterion cannot explain punishment behavior, as punishment strictly decreases own payoff and social surplus while weakly decreasing the minimum payoff in the institution (see Appendix 1 in their paper for both versions of the model).

A.1.1 Utility under inequity aversion due to Fehr and Schmidt (1999)

Let $\pi = (\pi_1, \ldots, \pi_n)$ denote the vector of monetary payoffs. Then, player i's utility is given by

$$u_i(\pi) = \pi_i - \alpha_i \left(\frac{1}{n-1}\right) \sum_{j \neq i} \Big[\max\{x_j - x_i, 0\} \Big] - \beta_i \left(\frac{1}{n-1}\right) \sum_{j \neq i} \Big[\max\{x_i - x_j, 0\} \Big] \quad (7)$$

with the conditions that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i \leq 1$. In the utility function, α_i captures the decrease in utility due to disadvantageous inequality, while β_i captures the decrease in utility due to advantageous inequality.

Proposition 1. *Without loss of generality, order the values of α_i such that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n$. Suppose there exists $q \in \{1, 2, \ldots, n\}$ such that preferences satisfy $(m/n) + \beta_j \geq 1$ and*

$$c < nr\alpha_j \quad (8)$$

for all $j \geq q$ and that $\alpha_j = \beta_j = 0$ for the remaining players, where $j < q$. In the public goods game with centralized punishment, all strategy profiles satisfying the following three properties constitute subgame perfect equilibria:

1. *Voting results in one of the players $j \geq q$ being elected as central authority with certainty. (Note that no specific member needs to be elected; moreover, ties are acceptable as long as $j \geq q$ for all candidates tied for the largest number of votes.)*
2. *During the contribution stage, $g_i = g \in [0, e_1]$ for all i.*

3. *If one of the players contributes $g_i < g$, the central authority assigns $(g - g_i)/r$ punishment points to player i.*

4. *Off the equilibrium path, any player $j < q$ elected to be the central authority will not punish because $\alpha_j = \beta_j = 0$. Using backwards induction, this lack of punishment results in zero contributions at the contribution stage.*

Proof. Suppose that one of the players contributes $g_i < g$ during the contribution stage. Let P denote the punishment assigned by the central authority to player i. Monetary payoffs are

$$\pi_i = y - g_i + \left(\frac{m}{n}\right)[(n-1)g + g_i] - rP - \left(\frac{c}{n}\right)P \tag{9}$$

for player i and

$$\pi_j = y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] - \left(\frac{c}{n}\right)P \tag{10}$$

for all $j \neq i$.

Notice, importantly, that player i's monetary payoff is reduced by both the received punishment and by her share of the cost of punishment, which is shared equally by all group members. We propose that the value of P that equalizes final payoffs will constitute an equilibrium.

$$\begin{aligned}
\pi_i &= \pi_j \\
y - g_i + \left(\frac{m}{n}\right)[(n-1)g + g_i] - rP - \left(\frac{c}{n}\right)P &= y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] - \left(\frac{c}{n}\right)P \\
-g_i - rP - \left(\frac{c}{n}\right)P &= -g - \left(\frac{c}{n}\right)P \\
g - g_i &= rP \\
P &= \frac{g - g_i}{r}.
\end{aligned}$$

While it should be clear that π_i is less than the equilibrium payoff, we include the algebra here for completeness. The monetary payoff from equilibrium is $\pi = y + (m-1)g$.

$$\begin{aligned}
\pi = y + (m-1)g &> y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] \\
&> y - g_+ \left(\frac{m}{n}\right)[(n-1)g + g_i] - \left(\frac{c}{n}\right)P \\
&= \pi_j \\
&= \pi_i.
\end{aligned}$$

Therefore, player i has no incentive to deviate, conditional on the punishment threat being credible.

Suppose the central authority reduces P by ε. The central authority's monetary payoff increases by $(c/n)\varepsilon$, as does the payoff of all other institution members (including player i). Unlike peer punishment, there is no inequity between the central authority and the other full contributors, as the cost of punishment is shared equally by all institution members. Player i's monetary payoff increases by $r\varepsilon$ from the reduction in received punishment and by $(c/n)\varepsilon$ from the reduction in the cost of assigned punishment. Therefore, the central authority, player j, suffers a non-monetary reduction in utility due to disadvantageous inequality in the amount of $\alpha_j r\varepsilon$. Thus, if $\alpha_j r\varepsilon > (c/n)\varepsilon$, the punishment threat is credible and the central authority prefers not to deviate from the proposed equilibrium. Since the ε term is common, the requirement reduces to $c < (nr\alpha_j)$, which is the condition stated in the proposition.

Notice also that the central authority does not have an incentive to punish player i beyond the point of equal payoffs. Any excess punishment causes a reduction in the central authority's utility by decreasing the monetary payoff due to increased costs of punishment and by increasing advantageous inequality with respect to player i.

Finally, we need to demonstrate that the central authority will not deviate in the contribution stage. The argument is the same here as in Fehr and Schmidt (1999)'s proof of their Proposition 5. The central authority can reduce her contribution to the public good by $\varepsilon > 0$ and increase her material payoff by $(1 - (m/n))\varepsilon$. Doing so creates advantageous inequality of ε relative to each of the other institution members, causing an overall decrease in utility of $\beta_i(1/(n-1))(n-1)\varepsilon = \beta_i\varepsilon$. The central authority will only deviate if $(1 - (m/n))\varepsilon > \beta_i\varepsilon$ or, equivalently, $(m/n) + \beta_i < 1$. This last condition is ruled out by assumption in the proposition. Therefore, the central authority will never deviate in the contribution stage.

The condition on voting is trivial.

□

A.2 Social Value Orientation

The Social Value Orientation (SVO) measure of Murphy et al. (2011) consists of six allocation decisions between oneself and one other anonymous individual. The SVO scale can be used to classify individuals as (1) altruistic, (2) prosocial, (3) individualistic, or (4) competitive.[12] The decision-making criteria for the four social preference types are

[12]The 6-item SVO scale cannot distinguish between prosocial individuals who are efficiency maximizers (Type 2.a) and prosocial individuals who are inequity averse (Type 2.b); an additional 9 items are provided by (Murphy et al., 2011) to distinguish between these two subtypes but are not relevant for the other classifications. Therefore, we rely on the 6-item measure as it provides a useful measure for all subjects. Moreover, distinguishing between efficiency-maximization and inequity aversion is not necessary or useful, as groups converge to the Pareto-dominant cooperative equilibrium of full contributions, which

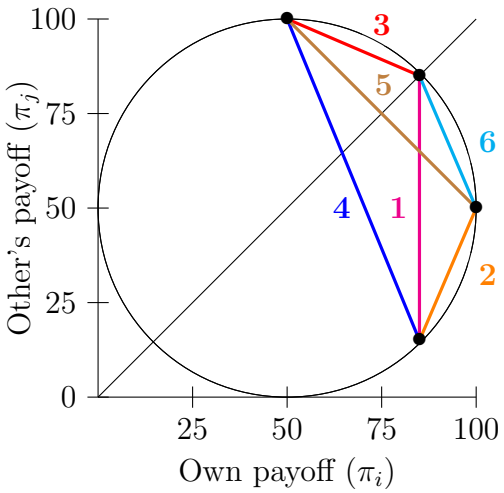


Figure 12: Social Value Orientation (SVO) allocation decisions. Each numbered line represents the set of options for one decision. Options consist of nine equally spaced allocations along the line. *Source:* Adapted from Murphy et al. (2011).

Type		Condition	Intuition
1	Altruistic	$\max\{\pi_j\}$	Maximize other's payoff
2.a	Prosocial (efficiency)	$\max\{\pi_i + \pi_j\}$	Maximize total payoff
2.b	Prosocial (inequity averse)	$\min\{\pi_i - \pi_j\}$	Minimize relative payoff
3	Individualistic	$\max\{\pi_i\}$	Maximize own payoff
4	Competitive	$\max\{\pi_i - \pi_j\}$	Maximize relative payoff

Table 5: Decision criteria for Social Value Orientation types.

given in Table 5.

The six allocation decisions are displayed by the numbered lines in Figure 12.[13] The SVO score is measured in degrees and is given by

$$SVO_i = \arctan\left(\frac{\overline{\pi_j} - 50}{\overline{\pi_i} - 50}\right) \tag{11}$$

where $\overline{\pi_i}$ ($\overline{\pi_j}$) is the average amount allocated to oneself (other person) over the six allocation decisions. Higher SVO scores indicate stronger social preferences.

is consistent with both subtypes.

[13]For example, line 2 in Figure 12 represents choosing among 9 equally spaced allocations ranging from (100,50) to (85,15), where the first entry is payoff to self (π_i) and the second entry is payoff to other (π_j). An individualistic person maximizes own payoff and would choose (100,50). A competitive person maximizes the relative difference between her own payoff and her counterpart's payoff; such a person would choose (85,15) because it yields the maximal relative payoff difference of 70.