Self-Organization for Collective Action: An Experimental Study of Voting on Formal, Informal, and No Sanction Regimes*

by Thomas Markussen[a], Louis Putterman[b] and Jean-Robert Tyran[c]

Abstract. Entrusting the power to punish to a central authority is a hallmark of civilization. We study a collective action dilemma in which self-interest should produce a sub-optimal outcome absent sanctions for non-cooperation. We then test experimentally whether subjects make the theoretically optimal choice of a formal sanction scheme that costs less than the surplus it makes possible, or instead opt for the use of informal sanctions or no sanctions. Most groups adopt formal sanctions when they are of deterrent magnitude and cost a small fraction (10%) of the potential surplus. Contrary to the standard theoretical prediction, however, most groups choose informal sanctions when formal sanctions are more costly (40% of the surplus). Being adopted by voting appears to enhance the efficiency of both informal sanctions and non-deterrent formal sanctions.

Keywords: formal sanctions, informal sanctions, experiment, voting, cooperation, punishment.

JEL Codes: C92, C91, D03, D71, H41.

---

* We wish to thank the Danish research council (FSE) for financial support under project title "Cooperation and Institutions," and Florian Spitzer for excellent research assistance.

[a] Department of Economics, University of Copenhagen, Denmark.

[b] Department of Economics, Brown University, Providence, RI, USA.

[c] Department of Economics, University of Vienna, Vienna, Austria and Department of Economics, University of Copenhagen, Denmark.

1. Introduction

One of the most fundamental puzzles of economic life is how people can cooperate given the preeminence of self-interest. In some domains, the problem is minimized because cooperation can take the form of mutually beneficial exchange. However, despite having such exchanges as its core feature, the modern market economy functions well only if supplied with an infrastructure of public goods including institutions (rule of law, a stable currency) and shared physical assets (roads, ports, etc.). The provision of these goods entails a collective action problem. Even in firms, which can interact with each other and with consumers primarily on the basis of mutually beneficial exchange, there is a need to elicit cooperation from employees. Small-scale collective action in communities and voluntary organizations also requires that the problem of cooperation be addressed.

In economic theories that assume strictly self-interested and perfectly rational agents having common knowledge of one another's type, two main solutions to the dilemma of cooperation are considered. First, cooperation is possible under conditions of indefinite repetition if individuals don't discount the future too steeply. Second, cooperation can be imposed by a central authority that uses the threat of penalties to enforce required contributions. In recent decades, economists have also devoted considerable attention to a third approach, that of voluntary cooperation (possibly aided by the opportunity of individuals to impose informal sanctions on one another), but appeal to such a solution in one-shot and finitely-repeated settings requires a departure from common knowledge of rational self-interest. Candidate points of departure include the potential presence of non-standard types (Kreps *et al*. 1982), altruism, warm glow (Palfrey and Prisbrey, 1997), inequality aversion (Fehr and Schmidt, 1999), and a preference for cooperation if others likewise cooperate (Fischbacher and Gächter, 2010).

This paper deals only with the second and third solutions, those of formal sanctions meted out by state intervention and voluntary cooperation, either with or without the backing of informal sanctions. We investigate behavior under and choices between formal sanctions, informal sanctions, and sanction-free regimes by conducting a laboratory experiment.

Recent years have seen a prodigious outpouring of experimental research on the impact of informal (i.e., peer-to-peer) sanctions, but considerably less attention has been devoted to formal sanctions. Correspondingly, the question of when informal sanctions are appropriate and when more formal rules and penalties are the better means of resolving conflicts between individual and group interests has not been addressed. This research gap is surprising given that the assignment of the right to punish to the state rather than to individuals is considered a distinguishing feature of civilized life. Further, the potential inadequacy of informal sanctions has been indicated by experiments suggesting that due to its costliness, frequent misdirection, and attempts to counter-punish, the opportunity to punish often fails to enhance cooperation and may even reduce efficiency.[1] Indeed, when offered a choice over whether to allow it, most subjects in experiments by Botelho *et al*. (2005) and Sutter *et al*. (2010) reject the option.[2]

While the studies just mentioned investigate how individuals respond to choices between an informal sanctions regime and a collective action setting without sanctions, no study of which we are aware has considered the choice of a sanctions regime when the option of a formal or centralized mechanism is also on the menu. Even small, voluntary organizations can choose between formal, vertical, or centralized enforcement methods, and informal, horizontal, or decentralized ones. For example, an organization can leave contributions strictly voluntary and use a mix of intrinsic identification with its goals, social pressure and social recognition to elicit varying levels of contribution from its supporters; but the same organization can choose to set required membership fees and use vertical enforcement methods, including the threat of removing membership privileges or imposing monetary penalties on those not meeting their obligations, to meet its revenue goals.[3]

---

[1] See Cinyabuguma, Page and Putterman (2006), Herrmann, Thöni and Gächter (2008), Denant-Boemont *et al*. (2007), Egas and Riedl (2008), and Nikiforakis (2008).

[2] When there are opportunities to reconsider the decision, however, the informal sanction institution becomes increasingly popular in experiments by Gürerk *et al*. (2006) and Ertan *et al*. (2009). We will return to this finding below.

[3] One early example of using penalties to enforce contributions to a non-state entity comes from 7th Century England, where a code quoted by Chaney (1963) stated: "Church dues shall be rendered at Martinmas. If anyone fails to do so, he shall forfeit sixty shillings and render twelve times the church dues (in addition)." Another setting in which the choice between formal and informal sanctions occurs is in promulgating rules of resource conservation, for example constraints on harvesting the

In choosing between formal and informal ways of addressing their free-riding problems, groups may consider many factors including the number of people served, their spatial distribution, the pattern of social links between them, and the excludability and rivalness of the goods and services to be provided. To keep the scope of our investigation manageable, we restrict our attention to pure public goods and abstract from spatial and other details. We focus our inquiry on three broadly relevant dimensions of the collective action and sanctions problem. These are (a) the targeting of sanctions, (b) the costliness of the sanctions regime, and (c) the motivational underpinnings needed to make the sanctions regime effective.

With regard to targeting, several studies demonstrate that a frequent problem with informal sanctions is that some of them are perversely or anti-socially directed not at free riders but at cooperators, for reasons among which the desire for retaliation seems to figure prominently (Cinyabuguma *et al*., 2006; Herrmann *et al*., 2008). Formal sanctions seem likely to be better targeted; for example in democratic groups, the majority is likely to favor a sanctions regime (if any) that is designed to overcome free-riding incentives.[4] Thus, the foundational British social and political philosophers Thomas Hobbes (1588 – 1679) and John Locke (1632 – 1704) argued that formal authority, or the state, emerges as a consensual response to the conflicts engendered by unbridled self-interest. In one passage, Hobbes (1996 [1651]) describes the state as a "visible power to keep [men and women] in awe, and tie them *by fear of punishment* to the performance of their covenants" (Chap. XVII, §. 1-2, p. 111, emphasis added). Locke (2005 [1689]) departs from Hobbes by entertaining the possibility that norms might be enforced (in what he terms "the state of nature") by informal sanctions, but he too concludes that sanctioning should be the job of the state due to the problem of retaliatory punishment:

---

forests around villages. In less developed countries, it may be better policy to encourage local people to make and enforce their own rules because governments lack adequate manpower and sufficiently inexpensive communications and transportation to monitor and punish violations, and in the absence of local "buy-in" the people on the ground may actively thwart or at least not actively help enforcement by the state (Ostrom, 2010).

[4] Although they study an informal sanctions context, the results in Ertan *et al*. (2009) are suggestive here.

> … every one in [the] state [of nature] being both judge and executioner of the law of nature, men being partial to themselves, passion and revenge is very apt to carry them too far, and with too much heat, in their own cases... (§. 125.)
>
> In the state of nature there often wants power to back and support the sentence when right, and to give it due execution. … such resistance many times makes the punishment dangerous, and frequently destructive, to those who attempt it. (§. 126.)
>
> … The inconveniencies that they are … exposed to, by the irregular and uncertain exercise of the power every man has of punishing the transgressions of others, make them take sanctuary under the established laws of government … It is this makes them so willingly give up every one his single power of punishing, to be exercised by such alone, as shall be appointed to it amongst them; and by such rules as the community, or those authorized by them to that purpose, shall agree on. (§. 127.)

The costliness of sanctions intersects with the targeting question in an interesting way. A formal sanctions mechanism so predictable as to render free riding clearly harmful to the individual may support higher compliance rates than an informal scheme with unpredictable punishment decisions by individuals. If effective enough, sanctions may rarely need to be imposed by the formal system because there may be few violations of the rule or norm in question. While this saves costs of sanctioning for both the punishing authority and the potential target of punishment, to have such a mechanism convincingly in place may require the incurring of fixed costs for the building, staffing and maintenance of tax collection offices, police departments, courts, prisons, etc. Whether the greater certainty and thus less frequent active use of a formal scheme induced by such high 'up front' costs is preferable to the potential uncertainties and presumably more frequent misdirection of punishment in an informal enforcement system can be expected to hinge on a variety of factors, including the inclination of individuals to bear the private cost of punishment and the frequencies of pro- and anti-social orientations among prospective informal punishers.

With regard to motivation, the choice between formal and informal sanctions is clear-cut from the standpoint of standard theory but complex from a more behavioral perspective. If individuals are assumed to be perfectly rational and self-interested, then allowing them to engage in costly punishment can have no effect outside of an indefinitely repeated context, since it is never rational to throw the "good money" of costly punishment after the "bad money" you have lost due to the targeted individual's free riding. In contrast, it is quite rational to vote for an effective formal sanctions scheme so long as what it costs to shift a group from an equilibrium of universal free riding to one of socially optimal cooperation is less than the aggregate benefit to the participants and an acceptable sharing

of costs can be worked out. In practice, though, numerous individuals make voluntary contributions and engage in costly punishment, indicating that motivation is more complex than implied by received theory. If forthcoming informal sanctions are sufficiently predictable and well-targeted, they might well be preferred to formal ones to which a high price tag is attached. That the presence of a collective action problem need not in and of itself make a centralized solution optimal, and that voluntary cooperation is sometimes an alternative to state authority, are themes more reminiscent of the 20th century view of Buchanan and Tullock (1962) and the 21st century view of Ostrom (2010) than of the earlier views of Hobbes and Locke.[5]

We conduct experiments in which subjects facing a linear, finitely repeated voluntary contribution problem are given the opportunity to vote on whether to use informal sanctions, formal sanctions, or no sanctions. We use the voluntary contributions mechanism because it is easy for subjects to understand, captures essential features of the problem of social cooperation, and has been studied extensively in variants with and without informal sanctions.[6] To eliminate strategic considerations from the voting process, we ask our subjects to choose between pairs of options before each of the experiment's six phases, [7] deriving what conclusions we can about preferences over the full set of three options from the votes on paired alternatives.

We study four treatments resulting from the 2x2 crossing of two dimensions of variation. In two treatments, formal sanctions carry a high fixed cost, whereas in two others, they carry a low cost. These represent 40% and 10%, respectively, of the potential gain

---

[5] Buchanan and Tullock write that "The existence of external effects of private behavior is neither a necessary nor a sufficient condition for an activity to be placed in the realm of collective choice (1962, p. 57)." Ostrom (2010) provides several examples that "challenge the presumption that governments always do a better job than users in organizing and protecting important resources" (p. 641) and asserts that "the earlier theories of rational, but helpless, individuals who are trapped in social dilemmas are not supported by a large number of studies using diverse methods" (p. 659).

[6] A more realistic model of the public goods problem would have an interior equilibrium in which some but not all resources would be used for public goods provision. This has been introduced into a few experiments such as Andreoni (1993) and Hichri (2004), but we judged the cost in terms of extra complexity to be unwarranted given our particular aims.

[7] With three options, participants may sometimes choose to vote for their second-most preferred scheme in order to prevent the implementation of the one they favor least.

from cooperation relative to full free-riding. Either cost should be rationally taken up according to standard theory, but might not be if subjects behave more cooperatively than theoretically predicted in no sanction or informal sanction environments. The other dimension over which our treatments are varied is the size of the penalty in the formal sanctions regime. In two treatments, the penalty is large enough to fully deter free riding by a self-interested individual, while in the other two, it falls short of that threshold, and is thus theoretically "non-deterrent" in character (Tyran and Feld 2006). Non-deterrent sanctions are of interest because they are arguably more common than deterrent ones in real world settings.[8] Because the efficacy of sanction regimes may depend on the concurrence of those operating under them, we check for the effects of voting on their performance by introducing also exogenous sanctions treatments, to be detailed later.

We find, unsurprisingly, that a majority of subjects vote for and achieve high efficiency with low-cost and deterrent formal sanctions. As for informal sanctions, we find that most subjects are initially disinclined to allow them, and that their common drawback—perverse or anti-social punishment—is indeed present. But we also find, like previous studies that allow groups to revisit their choices (Gürerk *et al*., Ertan *et al*.), that using informal sanctions becomes increasingly popular as experience of the collective action dilemma increases.

Somewhat more surprisingly, we find that subjects tend to prefer informal over even deterrent formal sanctions when the latter are relatively costly, despite standard theory's prediction that formal sanctions would be preferred. Indeed, voting outcomes depend on the cost at least as much if not more than on the deterrence level of the available formal sanctions, contrary to the prediction from standard theory. Moreover, even in the low-cost deterrent formal sanctions treatment, a substantial number of groups select informal over formal sanctions in their last vote. These outcomes are largely explicable by the fact that many individuals do incur the cost of punishing, contrary to standard theory but consistent with theories of negative reciprocity (Fehr and Gächter, 2000a) and altruistic

[8] One reason is that the penalty required to achieve deterrence may be considered to violate social standards of reasonableness, in part because of the possibility that violation occurred due to error or ignorance or that a rule-complying individual is wrongly penalized. A non-deterrent sanction may nonetheless deter most rule violation when it expresses a norm that citizens internalize or when being caught violating the rule brings informal as well as formal penalties, e.g. social disapproval.

punishment (Fehr and Gächter, 2002), and by the fact that punishment is mainly directed at free riders.

Finally, we find evidence that the efficacy of both informal sanctions and non-deterrent formal sanctions is enhanced when their use is selected by group vote—as seen from a comparison to a treatment in which subjects interact under an informal sanctions scheme assigned to them by the experimenter.[9] And we find that preference for formal over either no sanctions or informal sanctions varies with personal characteristics and experience. In particular, more reflective subjects are more likely to vote for informal sanctions and perverse punishers are less likely to vote for either sanction scheme when the alternative is no sanctions, while recipients of perverse punishment are more likely to vote for formal sanctions when the alternative is informal ones.

The rest of the paper proceeds as follows. Section 2 reviews the literature on formal and informal sanctions, explains how such regimes are expected to work based on standard economic theory, what the experimental and field evidence says, and what behavioral approaches have been put forth to explain it. Section 3 describes our experimental design. Section 4 presents the experimental results. Section 5 summarizes and provides concluding remarks.

2. Theory and literature

Consider a group of *N* individuals who obtain utility from consuming a pure public good and a private good. We assume for simplicity that each individual has the same endowment and the same utility or payoff function. Each unit allocated to providing the public good generates aggregate utility for the group members strictly exceeding the loss of utility to the individual donor, so that all are better off if all of their endowments are fully allocated to it.

[9] That voting choice improves the effectiveness of informal sanctions relates to a core issue in the informal sanctions literature: that the presence of norms or of coordination devices of some kind may be a pre-requisite to the effectiveness of informal sanctions. Herrmann *et al*. (2008) provide striking evidence of the importance of social norms by showing that these vary among societies. Ostrom *et al*. (1992) and Janssen *et al*. (2010) indicate that communication is a highly effective way to coordinate the use of informal sanctions, thereby greatly increasing its efficiency. Boyd *et al*. (2010) posit that coordination of punishment played a critical role in the early evolution of a propensity to punish.

However, an individual's own payoff is higher the more units she devotes to her private good. Each individual's payoff is thus given by

$$\pi_i = \left(E_i - C_i\right) + m\sum_{all\ j} C_j \qquad (1)$$

where E_i is the individual's endowment, C_i is her contribution to the public good, m is the marginal per capita return from contributing to the public good, hereafter MPCR, and j includes i. We impose $1/N < m < N$ so that the socially optimal payoff per person, mNE_i, exceeds the payoff when each individually optimizes, E_i. In some conditions, with approval of $V > N/2$ members, a formal sanction scheme can be put in place under which an individual incurs a sanction of s units for each unit she allocates to the private rather than the public good. Operating the formal sanction scheme requires payment of a fixed cost of $c < (mNE_i - E_i) = \mathcal{P}$. The right hand side of the inequality is the difference between an individual's earnings under full cooperation with c of zero and earnings under individual optimization. (We use notation $\mathcal{P}$ for "cooperation premium.") When the scheme is adopted, then, i's payoff becomes

$$\pi_i = \left(E_i - C_i\right)\left(1 - s\right) + m\sum_{all\ j} C_j - c \qquad (2)$$

where again j includes i. When $s > (1 - m)$, we say that we have a "deterrent" formal sanction, because the presence of the penalty deters free-riding by making it privately rational to contribute all of one's endowment to the public good. Accordingly, presence of the penalty changes equilibrium play among rational, self-interested agents from that in which $C_i = 0$, all i, to one in which $C_i = E_i$, all i. Since $mNE_i - c > E_i$ (equivalently $\mathcal{P} > c$), each enjoys a higher payoff with the scheme than without it, and it is therefore a dominant strategy to vote for the scheme when expecting to be a pivotal voter. It is a weakly dominant strategy to do so when unsure of others' votes.[10]

Now suppose that rather than an opportunity to enact a formal sanctions scheme, what is on offer is a contrasting scheme under which each group member has the opportunity to impose costly punishment on other group members after seeing how much

[10] The trembling hand perfection concept of Selten (1975) can also be invoked to justify the assumption of voting for the scheme.

they have contributed to the public good. Specifically, under an informal sanction scheme, any individual *i* can impose a cost σ on any other group member *j* at a cost to herself of one unit. Letting R_{ij} be the number of units of punishment (earnings reduction) member *i* decides to impose on member *j*, σR_{ij} indicates the loss that *j* incurs, so the payoff of an individual *i* under the informal sanctions scheme is

$$\pi_i = \left(E_i - C_i\right) + m\sum_{all\ j} C_j - \sum R_{ij} - \sigma \sum R_{ji} \qquad (3)$$

It is easy to see that in a one-time interaction, a rational individual *i* seeking to maximize her payoff will choose $R_{ij} = 0$, all *j*, even though *i* gets to learn C_j before choosing R_{ij} and can condition her punishment of *j* on it. By backward induction, the same logic extends to a finitely repeated interaction, if one assumes common knowledge that all group members are rational payoff-maximizers. Standard theory accordingly predicts no punishment, equilibrium behavior will thus have $C_i = 0$, all *i,* and payoffs will be E_i per person, identical to those when no sanctions are available. Rational individuals are therefore completely indifferent if offered the opportunity to vote on whether or not to allow informal sanctions, as a result of which the probability that an informal sanctions scheme will be selected by vote is in theory 0.5.

Finally, consider a third and last type of sanction scheme, a formal sanction of the kind described by equation (2) but where $0 < s < (1 - m)$. Following Tyran and Feld (2006), we call this a non-deterrent sanction because a rational payoff-maximizing individual selects $C_i = 0$ despite its presence. But with $0 < s < (1 - m)$, implementing the scheme changes *i*'s payoff to $E_i - s < E_i$. Accordingly, theory predicts that if offered the chance to vote on whether to implement a non-deterrent sanction scheme, individuals who perceive any chance of being pivotal will vote against it.

Summing up, the theory of rational agents who maximize own payoffs and have common knowledge that all are of the same type predicts (a) full free riding ($C_i = 0$) in the absence of sanctions, under non-deterrent formal sanctions, or in the presence of informal sanctions opportunities, (b) full contribution ($C_i = E_i$) under deterrent formal sanctions, (c) no punishment ($R_{ij} = 0$) if informal sanctions are available, (d) acceptance in a vote for deterrent formal sanctions when $c < \mathcal{P}$ and rejection of non-deterrent formal sanctions at

any $c > 0$, if there is a chance of the vote being pivotal, and (e) indifference between informal sanctions and no sanctions at all.

Some of these predictions have been tested experimentally.[11] These experiments provide considerable evidence that neither prediction (a)—zero contributions without deterrent sanctions—nor prediction (c)—no informal punishing in one-shot or finitely repeated environments—hold in practice. With respect to (a), the average amount contributed to a public good in voluntary contribution mechanism experiments is about half of the endowment in the first period of repeated play, with a typical contribution average of around a quarter of the endowment over ten rounds of repeated play. With respect to (c), the literature shows that costly punishment is common, and that the majority of the punishment is directed at lower contributors to the public good. Typically, the presence of informal sanction opportunities increases contributions, and often it reverses the otherwise observed trend of decay.

The fact that experiments with informal sanctions are usually associated with the elicitation of considerable punishment and with more sustained contributions to the public good should *not* lead us to predict unambiguously that self-interested subjects would vote for an informal sanctions scheme rather than play a voluntary contribution game without any scheme. While contributions to the public good are larger with sanctions and while *gross* earnings are larger with larger contributions, *net* earnings are often smaller once the costs to both the punisher and the target of punishment in (3) are taken into account (see, for example, Fehr and Gächter, 2000b, 2002, Bochet *et al*., 2006, Sefton *et al*., 2007). Because punishment is unpredictable and is sometimes aimed at high rather than low contributors, individuals might also oppose informal sanctions for other reasons, such as risk-aversion or concern about fairness. These factors might help to explain why in the existing experiments in which subjects were offered a choice between playing a voluntary contribution game without and one with informal sanctions, most subjects initially opposed the sanction regime.

---

[11] Surveys of experimental environments with no sanctions include Davis and Holt (1993) and Ledyard (1995). The survey portion of Gächter and Herrmann (2009) and Chaudhuri (forthcoming) expand the terrain to include experiments with informal sanctions.

Repeated observations in the experimental literature have suggested preference-based explanations of why the predictions of standard theory do not prevail. They include that an unanticipated restart of play in a treatment without sanctions generates a jump in contributions (Andreoni, 1988), that contributions tend to be higher in repeated play with a group of constant composition than in a randomly changing or "perfect stranger" group, and that average contribution increases with the MPCR although in theory it should remain zero for any MPCR < 1. Among the suggested departures from simple payoff maximization that have been invoked to explain these observations are altruism, warm glow, social welfare (aggregate payoff) preference, inequality aversion, and reciprocity or conditional willingness to cooperate. A conditional cooperator, for example, assigns a higher subjective payoff to contributing than to not doing so, when others contribute. This preference, akin to what Sen (1967) dubbed "assurance game preferences" in a prisoners' dilemma context, might be associated with a tendency towards positive reciprocity, while the willingness to incur a cost to punish a free rider can be seen as an instance of negative reciprocity (Fehr and Gächter, 2000a).

One way to explain responsiveness to changes in MPCR and to non-deterrent formal sanctions is to begin with the concepts of altruism and warm glow (Palfrey and Prisbrey, 1997), in which the individual receives a positive utility payoff from contributing to the public good.[12] If substantial numbers of subjects experience altruistic or warm glow benefits from contributing but if the size of the benefits differs from one individual to another, we could expect the number contributing to the public good to increase as the MPCR rises and reduces the private monetary cost per unit contributed. A non-deterrent sanction has the same effect of lowering the opportunity cost of contributing to the public good as does raising the MPCR, so a sufficiently high non-deterrent sanction might also lead some having altruistic or warm glow preferences to contribute to the public good. Expectation of such contributions could in turn lead individuals having conditionally cooperative preferences to add their own contributions (see Thöni *et al*. 2009). Such a combination of willingness by some to contribute when net private cost remains positive but is reduced, with willingness by the same or other individuals to contribute conditional on others doing so, might explain why Tyran and Feld (2006) found that having a non-

[12] The two motivations differ in the first, the payoff varies directly with the size of the gain to others while in the second, a fixed payoff arises simply from "doing the right thing."

deterrent sanction significantly increased contributions to a public good when the scheme was chosen by a majority of the group members.[13] Those authors conjecture that group members treated the voting outcome as a signal of an intention to cooperate (or in the terms just discussed, willingness to contribute when private cost falls to the level associated with the non-deterrent sanction), a plausible conjecture given that voting for a non-deterrent sanction is clearly not optimal if one expects complete non-contribution. Thus, we may also see some voting for costly non-deterrent sanctions in our own experiment, contrary to the more standard prediction discussed earlier.

With respect to deterrent formal sanctions, the behavioral phenomena noted provide no reason to revise the standard theory prediction that individuals will favor them over no sanctions when their cost is modest. However, if that cost constitutes a significant fraction of the earnings difference between a zero contributions and a full cooperation equilibrium—that is, if $c/\mathcal{P}$ is large—then payoff-maximizing individuals may prefer to operate without the scheme since behaviorally realistic expectations regarding contributions are substantially higher than the zero contribution prediction of standard theory. And voting against formal sanctions might be read as a signal of intent to cooperate voluntarily and thereby attain at least some of $\mathcal{P}$ without incurring cost *c*.

Since informal sanctions have a mixed record of sometimes improving and sometimes worsening efficiency relative to a no sanctions condition, predicting how subjects will vote about allowing them is difficult. Judging by the existing experiments mentioned earlier, we may anticipate some initial reluctance to let group members punish each other. At the same time, the experiments of Gürerk *et al*. (2006) and Ertan *et al*. (2009), which let subjects revisit this decision repeatedly, show a growing preference for an informal over a no sanctions condition. This leads us to conjecture that after experiencing the frustration of free riding (and possibly, in some of our designs, after seeing the corrective powers of formal sanctions) our subjects might increasingly choose to allow

[13] A more recent paper with voting on non-deterrent formal sanctions is Kamei (2010).

informal sanctions, especially if punishment is well targeted and thus increases earnings as well as contributions.[14]

Although the present paper is the first, to our knowledge, to study the choice between formal and informal sanctions, simultaneous research by Kamei, Putterman and Tyran (2011) studies the same issue in a different but complementary fashion.[15] Among other things, the experiments differ in that in Kamei *et al*., subjects must determine the details of a formal sanction scheme after deciding whether or not to use one. By making the dimensions of the sanctions schemes on which subjects are voting exogenous, our own paper's design permits the testing of more straightforward voting predictions, without strategic complications. With different subject pools and several other important differences of design, each of the two papers provides an important robustness check with respect to the other, and each suggests ways that some of the other's findings can be extended. We discuss and compare the results of Kamei *et al*. with our own findings in the conclusion of our paper.

3. **Experimental design**

Basic design

Our public goods experiment with endogenous institutions entails play under three conditions—no sanctions (NS), formal sanctions (FS) and informal sanctions (IS)—in four treatments distinguished by sanctions level and cost when FS is adopted, plus two exogenous treatments used to test for an endogeneity effect in the most commonly observed IS and non-deterrent FS conditions. In all treatments, participants are divided into groups of $N = 5$ members that remain fixed ("partner matching") throughout a set of

---

[14] The findings of Herrmann *et al*. (2008) suggest that the "if" in the last sentence is an important one, but those findings also suggest that our Danish subject pool will fall within the better-performing side of their cross-cultural behavioral spectrum.

[15] Rockenbach and Wolff (2009) might also be mentioned, but the set of potential mechanisms available to their subjects is much larger than in the experiments we discuss here, and it is up to their subjects to dream up their own mechanisms rather than select from a menu of options offered by the experimenters. Interestingly, their subjects' almost never proposed what they refer to as "peer-to-peer punishment," a fact that might be interpreted as further evidence of initial reluctance to allow informal sanctions.

interactions lasting 28 periods, divided into seven phases of 4 periods each. Every period, each participant receives an endowment of E_i = 20 points of experimental currency. He or she decides how much of this endowment to allocate to a public good (referred to in the instructions as the "group account"), and how much to keep for him or herself (referred to as an allocation to a "private account"). The amount in the group account is scaled up by a factor of two and divided equally among all group members, thus m = 0.4. We refer to this standard voluntary contributions mechanism as the "No Sanctions" (NS) regime. In it, the individually optimal strategy is to contribute nothing to the public good, whereas the socially optimal solution has every group member contributing his or her full endowment.

As described in the previous sections, we give groups the opportunity to introduce two different types of sanctioning regimes that may contribute to solving this collective action problem, namely formal and informal sanctions.

Under *Informal Sanctions* (IS), participants observe at the end of the contribution stage of each period what fellow group members contributed to the public good.[16] They then have the opportunity to reduce the earnings of other group members, at a cost to themselves. Subjects learn the amount of punishment they receive, but not who gave it or how much punishment others receive in total.

Under *Formal Sanctions* (FS), allocations to the private account are penalized at a fixed rate s per point and participants pay a fixed cost c per period to have the scheme in place. Penalties are lost not only to the penalized subject but also to the group, and are not otherwise redistributed. The values of s and c are fixed for a given treatment but vary across the four main treatments, as detailed below.

Groups choose whether to play with NS, IS or FS through a series of votes. In each vote, only two institutions are available for choice. Voting is compulsory, simultaneous, and free, and each subject must vote for one of the institutions available (i.e. it is not possible to submit a neutral vote). The institution receiving the majority of votes in the group is implemented, with subjects learning what scheme was chosen but not the specific number of votes for it.

[Figure 1 about here]

[16] To ensure comparability across different regimes, subjects were always informed about the contributions of each other group member, even when informal sanctions were not used. In all treatments, information about the contributions of others is presented in a random order to preclude individual reputation formation.

Figure 1 shows the sequential structure of the experiment. We first hand out instructions for the No Sanctions regime, read aloud a brief summary, make sure that all subjects correctly answer a set of control questions testing their comprehension, and privately answer any questions. All groups then play four periods under this exogenously imposed regime. Then, a second set of instructions is distributed, explaining the rules of formal and informal sanctions, the voting system, and the fact that there will be six votes, each governing four periods of play.[17] These instructions also are accompanied by brief oral instructions, control questions, and answering of any questions raised by the subjects. At the beginning of Phase 2, each group chooses between informal and no sanctions, then at the start of Phase 3, between formal and no sanctions, and at the beginning of Phase 4 , between formal and informal sanctions, with the cycle repeated in phases 5 – 7. Although the order in which institutions are introduced might potentially affect outcomes, we chose to keep this order fixed while focusing on the variation of s and c.[18] Also, the repetition of the voting cycle reduces potential problems related to order effects, since all subjects have experienced at least two regimes when the second voting cycle is reached, and many will have experienced all three.

We use neutral language, avoiding terms such as "public good", "contribute", "free rider" or "punishment." (Instructions are included in the appendix.) At the end of the experiment, subjects answer a set of questions measuring political attitudes, risk preferences, and gender. They also take a three-question "Cognitive Reflection Test" (hereafter CRT; see Frederick, 2005), included to obtain a tested and relatively quick measure of cognitive ability, and they select a political self-identification on a five point scale ranging from very liberal to very conservative.[19]

---

[17] The reason for handing out two separate sets of instructions, and for having the initial phase with the No Sanctions regime exogenously imposed on all groups, is that it is considerably easier for participants to understand the rules of formal and informal sanctions once they have familiarized themselves with the No Sanctions version of the public goods experiment.

[18] A reason for introducing formal sanctions last is that in terms of human history, such sanctions regimes are arguably the last to arrive on the scene, in keeping with their association with "civilization".

[19] The CRT task was not incentivized.

Payoffs

Payoffs under each scheme follow equations (1) – (3) above, applying specific parameter values. With $E_i = 20$ and $m = 0.4$, π_i, under the No Sanctions (NS) regime is derived from (1) as:

$$\pi_i^{NS} = 20 - C_i + 0.4\sum_{j \in g} C_j \qquad (1')$$

Payoffs under the formal sanction scheme can be written as

$$\begin{aligned} \pi_i^{FS} &= (1-s)(20 - C_i) + 0.4\sum_{j \in g} C_j - c \\ &= 20(1-s) + (0.4 + s - 1)C_i + 0.4\sum_{j \neq i} C_j - c \end{aligned} \qquad (2')$$

where the sanction rate s and the scheme's cost c are free to differ between deterrent versus non-deterrent and cheap versus expensive sanctions treatments. Specifically, we implement four treatments, defined by the parameters of formal sanctions, according to the following table:

[Table 1 here]

With $s = 0.8$, formal sanctions are deterrent, while with $s = 0.4$ they are non-deterrent, since (2') implies that whether full contribution to the public good is or is not individually optimal depends on whether s is above or below 0.6. The two values that c takes, 2 and 8, represent 10% and 40%, respectively, of the hypothetical gains from full cooperation ($\mathcal{P}$), a difference we think substantial enough to justify distinguishing them as "cheap" versus "expensive." The interaction of the two dimensions yields the four treatments Deterrent Cheap (**DC**), Deterrent Expensive (**DE**), Non-deterrent Cheap (**NC**) and Non-deterrent Expensive (**NE**). The parameters of formal sanctions were fixed throughout each session of the experiment, with subjects being informed about these parameters in the instructions handed out after Phase 1.

In the informal sanctions scheme, for each reduction point allocated, the earnings of the *receiver* are reduced by four points, and the earnings of the *sender* are reduced by one point; hence σ of equation (3) equals 4.[20] Each subject is allowed to allocate at most 10

[20] Compared to some related experiments (e.g. Fehr and Gächter 2002, Egas and Riedl 2008), it is cheaper to reduce the earnings of other group members in our experiment, something Nikiforakis

reduction points to each other group member in each period. Also, reduction points *received* can never reduce a subject's earnings for the period to less than zero. However, reductions points *sent* must always be paid for, even if this leads to negative, total earnings for the period.[21] With these rules, earnings under IS are given by (3'), which modifies (3) using the values of E_i and σ and the just-stated limit on punishment losses, that is

$$\pi_i^{IS} = \max\left(0, 20 - C_i + 0.4\sum_{j\in g} C_j - 4\sum_{j\neq i} R_{ji}\right) - \sum_{j\neq i} R_{ij} \qquad (3')$$

where $\sum_{j\neq i} R_{ji}$ denotes the number of reduction points received by individual *i* from individuals *j*≠*i* and $\sum_{j\neq i} R_{ij}$ the number of reduction points sent by *i* to any *j*≠*i*

Predicted behaviors based on standard economic theory at the values here specified follow immediately from the logic discussed in Section 2.[22]

---

and Norman (2008) find contributes to punishment's ability to increase both contributions and earnings. The 1:4 cost ratio is used in Page *et al*. (2005) and Bochet *et al*. (2006) but seems somewhat less effective there, perhaps because the opportunity to purchase punishment in smaller increments of 0.25 experimental dollars reduced the amounts demanded. Even if cheap punishment increases efficiency, it is far from clear that it also increases the popularity of informal sanctions. Sutter *et al*. (2010) find that subjects were more likely to vote for informal sanctions with a cost ratio of 1:1 than for sanctions with a ratio of 1:3—suggesting a fear of entrusting anonymous others with lethal weapons.

[21] The maximum of 10 punishment points is common to a number of experiments in the literature and is rarely binding, being reached in less than 1% of punishment actions in our data (for cases in which punishment was purchased, its mean amount was 1.8 with standard deviation 1.4). We limited the effect of punishment so that period earnings could not become too negative, a matter of concern due to our desire to keep each period relatively independent in an accounting sense, and the difficulty of asking subjects to pay the experimenter. Similar to the previous constraint, this one was also rarely binding, the percentage of period-and-subject punishment events in which it provided protection amounting to 1.04%. Finally, we chose to have no exception to incurring the cost of punishing because we see costliness to the punisher as too important a feature of punishment to be waived. In the event, only 32 of the 2,280 subject-periods (1.4%) under IS saw a subject incur negative earnings for the period, and no subject's aggregate earnings from the session's main experiment approached 0 points (the lowest observed being 465 points, occurring in the **NE** treatment).

[22] Since the deterrent sanction $s = 0.8$ should be associated with full contributions and since the two values we implement for c (2 and 8) are both well below the threshold $c < \mathcal{P}$, theory predicts that participants will always vote for FS in both the **DC** and **DE** treatments. With non-deterrent sanctions

Exogenous treatments

To investigate whether efficiency achieved in the most commonly observed conditions of voted informal and non-deterrent formal sanctions is attributable to choice of those schemes by voting, we conduct parallel treatments in which subjects experience the same sequences of conditions of play, including identical cost parameters, but without ever voting on the conditions under which they will interact (see the right column of Table 1). Details of the treatments are given when we report the corresponding tests for effects of endogeneity.

4. Results

The experiment was conducted at the Centre for Experimental Economics, University of Copenhagen, between October and December, 2009. We conducted three experimental sessions for each treatment. The number of subjects per treatment varies from 60 to 70 due to no-shows by registered subjects (see Table 2). The number of groups per session varied from 3 to 6, with 5 as the most common number. In total, 260 subjects participated in the endogenous treatments, with a further 75 in the exogenous ones. Slightly over half of participants (51 percent) were freshmen economics students, about two months into their studies. The rest were from many different fields of study at the University of Copenhagen. 43 percent of participants were women. The experiment was conducted in a computer lab, using the software Z-tree (Fischbacher 2007). At the end of the experiment, each subject's earnings from all 28 periods were converted into money (1 point = 0,2 Danish kroner). Subjects earned on average 172 Danish kroner (about 33 USD). Each session lasted about one hour and 45 minutes.

Since our ultimate interest is in subjects' choices between sanction regimes as well as between sanction and sanction-free regimes, we begin our discussion of experimental behavior with a description of the voting outcomes. We then discuss contribution (and in IS also punishment) behaviors under each condition and treatment, and their earnings

---

(s = 0.4), on the other hand, zero contributions to the public good re-emerge as the dominant strategy for each individual, so predicted per period earning $\{20(1 - s) - c\}$ equal 10 (4) in treatment **NC** (**NE**), and subjects should thus vote against FS in the latter treatments. Given indifference between IS and NS according to standard theory, this implies preferences of FS > NS ~ IS in the **DC** and **DE** treatments and NS ~ IS > FS in the **NC** and **NE** treatments.

consequences. We use that information to get a sense of the efficiency of subjects' institutional choices before briefly discussing an individual-level analysis of voting using multivariate regressions. We end the section by comparing play under IS and non-deterrent FS conditions when determined endogenously vs. when imposed exogenously.

Voting

Figure 2 displays the voting outcomes by group in our four main treatments. In their first vote, the large majority (around 80%) of groups in every treatment rejected the option of allowing informal sanctions in favor of a no sanctions regime, in line with the previous literature. Second votes, between formal and no sanctions, vary considerably among treatments, with the proportion favoring sanctions being over 70% in **DC** and nearly 60% in **NC**, but under 20% in **DE** and under 10% in **NE.** This voting pattern doubly violates standard theory, by rejecting a deterrent sanction costing only 40% of the cooperation premium $\mathcal{P}$ in treatment **DE**, and by opting for a non-deterrent sanction with positive cost, in treatment **NC**. The pattern might make behavioral sense, however, if as in past experiments there is some cooperation in the absence of sanctions (which might render NS more efficient than FS in treatment **DE** but not in **DC**) and if non-deterrent sanctions have a substantial effect on contributions (making FS more efficient than NS in treatment **NC**). We'll show later that these past findings are indeed echoed by our subjects' behaviors, although on average the use of non-deterrent formal sanctions in treatment **NC** is not effective enough to cover their cost.

The third vote, immediately before Phase 4, pits informal sanctions, originally unpopular in all treatments, against formal ones, which were popular when cheap. Standard theory predicts a preference for FS when it is deterrent and for IS when FS is non-deterrent. The result we observe is that more than 80% of vote outcomes favor informal sanctions when formal ones are expensive (c = 8), a result that might, as in vote 2, reflect the desire to avoid the fixed cost of FS, with subjects presumably hoping that there will not be much voluntary punishing under the IS alternative. In the two cheap sanctions treatments (c = 2), majorities in more than 60% of groups vote for informal sanctions when formal sanctions are non-deterrent (**NC**), but a surprising 43% also try informal sanctions when formal ones are deterrent (**DC**).

The fourth vote begins the second cycle of voting, with all groups having experienced at least one sanctions regime and at least four periods without sanctions.[23] Against this backdrop, the second match-up of informal and no sanctions shows a considerable gain in popularity for informal sanctions, which go from adoption by 21% of groups (in the first vote) to 50% of groups (in the fourth vote) in the **DC** treatment, from 25% to 50% in **DE**, from 14% to 64% in **NC** and from 25% to 67% in **NE** (see again Fig. 2). The increasing popularity of IS with experience resembles the findings in Gürerk *et al*. (2006) and Ertan *et al*. (2009). Although in our experiment the path towards this outcome features choices involving the alternative of formal sanction that have no counterpart in any past study, exploratory regressions fail to show any relationship between groups' past experience of FS and their choice of IS in later votes.

The fifth vote, at the beginning of Phase 6, is the second contest between formal and no sanctions. It sees an 8 to 16% increase (in comparison with vote 2) in group choice of formal sanctions in all treatments except **NC**. In the latter, choice of formal sanctions drops (by some 14%), but they continue to be used by more than 40% of groups compared to less that 20% in **NE**, a difference evidently driven by the cost difference (see below).

Finally, the sixth vote is of particular interest as our second and last contest between formal and informal sanctions and the one in which subjects have the most prior experience to go by. In it, IS is preferred to FS by the majority of groups in all treatments except **DC,** in which FS is preferred by 57% of groups. There is no change (from the third vote) in voting shares in the **DC** and **NE** treatments, a small increase in the proportion of groups choosing IS in the **NC** treatment and a slightly larger decline in the proportion choosing IS in **DE**, where IS is nonetheless still preferred by about two-thirds of all groups.

Does an overall ranking of schemes emerge from these six votes in four treatments? If we rank conditions NS, IS and FS by number of groups favoring each in two-way match ups and if we focus on the last three votes, in which subjects are most experienced, we see the ordering IS > NS > FS in the two treatments with non-deterrent sanctions (**NC** and **NE**), IS ~ NS > FS in the **DE** treatment, and FS > IS > NS in the **DC** treatment. With experience, then, informal sanctions become at least as popular as formal ones except when the latter

[23] More specifically, 50% of subjects have experienced both formal and informal sanctions, 27% have only experienced informal sanctions and 23% only formal sanctions, as of this point.

are both deterrent and cheap, a finding that whets our appetites for information on how each institution performed. At least one of the sanction schemes is preferred by most to operating without sanctions except in the **DE** treatment, where IS and NS are preferred to FS but are each favored by the same number of groups.[24]

In addition to subjects' choices violating the standard theory prediction that deterrent formal sanctions should always be preferred to informal ones unless c exceeds $\mathscr{P}$, they also violate the prediction that the deterrent/non-deterrent distinction would be critical to voting whereas differences in c would have no impact as long as $c < \mathscr{P}$. To examine the impact of sanction cost versus deterrence more rigorously, we estimate a series of probit regressions at group level; see results in Table 3. Regression (1) indicates that cost of FS predominates and that the deterrence level of FS has an insignificant effect in Vote 2. In regression (2), we check whether having used IS in the phase preceding this vote made a difference, and find that it did not. In Vote 3, the first vote on FS versus IS, the deterrence level begins to show marginal significance, but cost of FS is still the only highly significant factor (see column (3)), and prior experience again lacks significant impact (regression (4)).

Regressions (5) to (8) indicate that the importance of the deterrence level as a determinant of voting rose over time, but that the costliness of FS remains important in the final votes. Regressions (6) and (8) also show that experience became important. In particular, regression (6) shows that groups that chose FS over IS in Phase 3 (IS over NS in Phase 5) are significantly more (less) likely to select FS over NS in Vote 5. Regression (8) shows a still more significant negative effect of past choice of IS on voting for FS in vote 6. This experience effect suggests that "to know IS is to love it," for our subjects.[25]

---

[24] Appendix table A.1 shows the exact vote shares corresponding to Figure 2 as well as corresponding shares at the individual level. Differences in individual and group vote shares are on the whole relatively small.

[25] We cannot rule out that some other factor caused both the past and the Vote 6 choice of informal sanctions.

Contributions and punishment

To explain the frequent preference for informal over formal sanctions observed in our data in contradiction of standard theory, we need to pay some attention to how subjects behaved under the competing sanction regimes or conditions.

Figure 3 displays average contributions by period in each of the four treatments and under each condition or regime. In each treatment, groups operating under a no sanctions regime—the standard VCM—display patterns familiar from other experiments: the average contribution begins around half of the endowment and eventually declines within a given phase, although there are strong restart effects if the scheme is chosen by voting in subsequent phases.[26] Deterrent formal sanctions lead to contributions between 80 and 100% of endowment, averaging 93% of the endowment in Phase 3 and 95% in phases 4, 6 and 7 of the **DC** and **DE** treatments. With anticipated full efficiency never quite achieved by deterrent FS, and with NS play showing average contribution levels considerably above zero, the range of costs over which a costly sanction would be profitable to impose is clearly smaller than the theoretical $c < \mathcal{P}$ (= 20) condition indicates.[27]

Non-deterrent sanctions, which in theory should have no power to induce contributions, are in fact associated with higher contributions than the NS regime during phases 3 and 6 of the **NC** and **NE** treatments, when each can be observed in some groups. In phases 4 and 7, when only some groups in **NC** treatment select them over IS, they also appear to generate higher contributions than are observed in phases under NS.[28]

---

[26] The relatively high contributions under NS in the first periods of later phases could also be due to some subjects reasoning that their group's vote for NS signaled a determination to achieve good results without costly sanctions, and to self-selection of more voluntarily cooperative groups into the condition.

[27] Consider, for instance, that the average contribution under NS typically averages more than 11 points in the phases of **DE** treatment in which play under that regime is possible. Since a subject earns 31 per period with all contributing 11 in NS and earns 39 per period with all contributing 95% of endowment or 19 in FS *before* deducting cost *c*, we see that $c = 8$ is in fact very close to the cut-off point for FS to be profitable under actual behaviors.

[28] Mann-Whitney tests show significant differences in average contributions of those groups using FS and those using NS in all but one case in phases 3 and 6 of the non-deterrent sanction treatments (see Appendix Table A.3). For those treatments we also conducted Wilcoxon signed-rank tests

Turning to informal sanctions, while the opportunity to impose them leaves unchanged the theoretical zero contribution prediction because rational players should not pay for punishment (assuming common knowledge and uniform type), Figure 3 shows that contributions under the IS condition resemble those under deterrent formal sanctions in treatments **DC** and **DE.** Contributions are similarly high and thus exceed those under non-deterrent FS in the **NC** treatment. In **NE** treatment, contributions under IS average roughly 70% of endowment, resembling average contributions under FS but clearly exceeding those in NS condition.[29] The empty circles near the bottom of Figure 3's quadrants show average sanctions given per subject when IS was in place. They make clear that sanctions were indeed used, although the amount is remarkably small in some phases (4, 5 and 7 of **DC**, 7 of **DE** and **NC**) in which contributions are nevertheless quite high.

A closer look at the pattern of punishment show how use of the IS scheme succeeded in raising contributions. As reported in Appendix table A.2, subjects used the opportunity to punish in an average of 14% of the 4 opportunities available to them each period, and 83% of subjects punished and 85% received punishment at least once. Subjects purchased an average of 1 punishment point per period under IS, and a subject who engaged in a positive amount of punishment in a period assigned 3.1 points of it on average. The average recipient of punishment was targeted with 3 points of punishment and thus lost 12 points (a large bite out of the theoretically predicted earnings of 20). Finally, the lion's share of punishment was relatively efficiently targeted, so that subjects tended to earn more by contributing more.[30]

---

comparing average contributions in all phases with FS and average contributions in the same groups in all phases in which they operated under NS (except phase 1 where institutions were not endogenously chosen). These within-group tests show contributions to be significantly higher with FS than with NS ($p < .01$) for **NC** groups but not for those in **NE**.

[29] Mann-Whitney tests show average contributions by group to be significantly higher for groups using IS than for those using NS in phases 2 and 5, apart from one treatment and phase. Average contributions do not differ significantly between groups using IS and those using FS in phases 4 and 7 of the **DC** and **DE** treatments, but are significantly higher for groups using IS than for those using FS in the corresponding phases of the **NC** treatment. This comparison cannot be made for **NE** since no groups chose FS in Phase 4 or 7. See Appendix Table A.3.

[30] This claim is supported by a series of GLS regressions for all IS-condition observations at individual level. See the discussion below Table A.2 in the Appendix.

The regressions of Table 4 explore further who punishes and when, using both GLS and Tobit specifications. As found in Önes and Putterman (2007), both the positive deviation of the targeted individual's contribution from the average contributed by other group members and the negative deviation from that average attain highly significant coefficients. We find that the magnitude of the negative deviation effect is about a third that of the positive deviation. Thus, most of the punishment is well targeted. We also check whether having been a recipient of either non-perverse or perverse punishment in the previous period increases one's likelihood of punishing.[31] The coefficients on the receipt of non-perverse punishment variable are all positive, large and highly significant in the three Tobit estimates, suggesting that revenge was a motive for punishing by punished low contributors. Receiving perverse punishment, in contrast, does not significantly affect punishing but the coefficients on this variable are consistently negative, hinting that "blind revenge" aimed at high contributing punishers may have deterred some from punishing free riders. The coefficients on the dummy variables for phases, although mainly insignificant, suggest a downward trend in the amount of punishment given, consistent with the pattern of circles in Figure 3. We defer for now discussion the coefficients on personal measures.[32]

---

[31] We classify an instance of punishment as perverse if the person targeted contributed more than the group's median amount for the period. Notice that the related concept of anti-social punishment is defined in Herrmann, Thöni and Gächter (2008) in terms of the relationship between the punisher's and the punished subject's contributions, whereas the definition of perverse punishment considers the contribution of the punished subject only. We find the latter approach more attractive when focusing on the effect upon the recipient of punishment in circumstances in which who gave the punishment is unknown to her. It is nevertheless reasonable to assume that the sets of perverse and anti-social punishers largely overlap.

[32] In Appendix tables A.5 and A.6, we include two tables investigating the determinants of contribution decisions by individuals under the IS, non-deterrent FS, and NS regimes. Highlights include (a) confirmation that first period contribution tends to be a good predictor of contribution throughout a session, (b) punished low contributors tend to significantly increase their next period contribution in IS condition, whereas punished high contributors tend to reduce their next period contribution, and (c) the previous period mean contribution of fellow group members is a significant positive predictor of own current contribution. (Results for variables referring to personal characteristics of the subjects are mentioned later.)

Efficiency and voting on sanction schemes

While we've seen that anticipation of both formal and informal sanctions helped to boost contributions to the public good, to know whether a particular sanctions scheme raised *efficiency*, we compare the costs of imposing and receiving sanctions with the social benefits of the higher contributions they helped to elicit. For a given group, sanction scheme, and phase, we define the gross gain in average earnings associated with the scheme as the group's average gross earnings during the phase in question (before deducting the cost of sanctions) minus the same group's average earnings under NS in Phase 1.[33] Dividing the result by average expenditure on sanctions per member and period gives us a measure of cost effectiveness of sanctions, defined as the average gain in earnings per point spent on sanctions.[34] The four panels of Figure 4 show this cost effectiveness indicator for IS and FS in the earlier and later phases, by treatment. The figure shows that on average, formal sanctions more than paid for themselves—thus proving preferable to NS—only in the **DC** treatment, whereas informal sanctions more than paid for themselves—proving preferable to NS—in the second halves of all treatments as well as in the first halves of the **DC** and **NC** treatments. Informal sanctions outperformed formal sanctions in direct comparisons in the second halves of all four treatments and in the first half of the **NC** treatment. The performance difference between the two kinds of sanction was relatively small in the three treatments in which efficiency of IS was less than that of FS in the first half.

We next ask how individuals' votes were influenced by their personal earnings experiences, then check the degree to which group voting outcomes favored the more efficient of the pairs of options offered, and whether success in that regard was improving over time. The first question can be answered only for individuals who had experienced both available schemes before a vote. Table 5 reports the marginal effects of individuals'

---

[33] While comparing contributions under IS or FS with those in other groups using NS in the same phase are possible in phases 2, 3, 5 and 6, using Phase 1 NS contributions provides a group-specific standard for both the IS and the FS comparisons that is also available for phases 4 and 7, in which there are no NS observations. The impact of IS or FS on contributions relative to NS may be somewhat understated by our method insofar as contributions in a no sanctions setting are typically lower in later than in earlier periods.

[34] In FS, sanctions expenditure includes administrative cost *c* and the cost associated with sanctions actually imposed. In IS, it includes cost both to punisher and to recipient of punishment. Average sanction cost per period by treatment is shown in Appendix Figure A.1.

past relative earnings in the two available schemes in the votes before each of phases 4 – 7 on voting using probit regressions with errors (with one indicated exception) clustered by group. With the exception of the first vote, where there are fewer observations and the significance level is 5%, all relative earnings terms attain significance at the 1% level with their signs indicating that subjects were more likely to vote for whichever scheme had previously given them higher earnings. Columns (3), (5) and (7) show specifications that include the ratio of the coefficients of variation of the individual's past earnings under the available schemes, to check whether variability of earnings was also a factor. The ratio terms obtain insignificant coefficients and their inclusion leaves the coefficients on relative average earnings largely unchanged.

While the regressions in Table 5 show that subjects generally tried to vote for the scheme that had given them higher earnings in the past, a great many instances of voting occurred without past experience of both offered schemes. It is interesting to investigate the extent to which group votes succeeded in selecting that scheme associated with higher earnings during the phase in question for groups in their treatment as a whole, and whether success in that respect increased with experience.[35] Those average earnings are displayed in Table 6. Overall, we find that 59.6% of groups made the "right" institutional choice in the first vote, 51.9% did so in the second vote, 48.1% in the third vote, 57.7% in the fourth vote, and 69.2% in both the fifth and the sixth vote. Since adjacent votes dealt with different pairs of options, the appropriate test of learning over time is to compare the first three to the last three votes. Here, there was a small upward trend in the proportion of groups choosing a scheme with higher earnings—53.2% doing so in the first three phases versus 65.4% in the last three.[36]

---

[35] Note that to simplify matters, we proceed on the assumption that by choosing scheme X in phase Z, a group would have achieved the average earnings observed under that scheme in that phase and treatment. Because no groups selected the NS alternative in Phases 4 and 7 in the NE treatment, we make the special assumption that earnings in that treatment under NS would have been the same as the average observed in Phases 3 and 6, respectively.

[36] Table 6 shows that the earnings difference among schemes is often only a few points or percent, and in 14 of the 22 cases in which at least some groups within a treatment used both schemes, the Mann-Whitney test value for difference in earnings by chosen scheme fails to achieve significance at the 10% level. Thus, assuming that all groups would do equally well with a given scheme and treating all "errors" of voting choice as being equally serious provides a crude indication, only, of the efficiency of voting choices and of learning over time.

Adding contributions and punishment to the explanation of voting

In additional regressions, we carried Table 5's analysis of individual voting decisions a step further, testing whether propensities to contribute and punish, and being the recipient of perverse punishment, may have affected subjects' votes. In particular, we included in these regressions the subject's period 1 contribution (arguably an indicator of propensity to cooperate), whether the subject has previously punished a contributor of more than the group median ("gave perverse punishment"), and whether the subject has herself been punished when contributing more than the group median ("received perverse punishment"). The estimates (shown in Appendix Table A.4) suggest that when deciding between a sanction scheme and having no sanctions, perverse punishers were significantly less likely to vote for either kind of sanctions.[37] When forced to choose between the two sanction schemes, having given perverse punishment no longer appears important, but having received it made a subject significantly more likely to favor FS over IS. That result fits neatly with the spirit of Locke's concerns, cited in the introduction. All coefficients on relative earnings remain highly significant and intuitively signed.

Effects of personal characteristics

Our experiment included a cognitive reflection test (CRT) because we conjectured that differences in cognitive sophistication might influence voting by allowing better inferences to be made about the relative advantages of the available options. We elicited a measure of political preference to test the conjecture that more conservative subjects might be more likely to oppose the use of sanctions owing to a distaste for coercion, as found by Putterman *et al*. (2010). To investigate their possible effects, we added the CRT and political preference variables along with a gender indicator to the regressions just discussed. The results (also shown in Appendix Table A.4) indicate that high CRT subjects were significantly more likely to favor informal sanctions over no sanctions in the vote preceding Phase 5, when IS was indeed the more efficient option. None of the other personal characteristics measures obtain significant coefficients, nor do a set of within-session experience variables added in some specifications.

---

[37] This result is consistent with the interpretation that the tendency to punish perversely is in part a manifestation of a generally uncooperative or anti-social orientation.

In the specifications not yet discussed in Table 4, we explore whether these same individual-specific measures may have influenced subjects' inclinations to incur the cost of punishing other group members. We also add to those regressions the individual's contribution in period 1 (played under exogenous NS condition in Phase 1), again as an indicator of propensity towards cooperation. Of the personal characteristics variables, the CRT score obtains significant negative coefficients in all four specification that include it, indicating that cognitively sophisticated subjects punished less than others under IS, all else being equal. General political preference obtains negative coefficients, highly significant in the Tobit regressions, indicating that more politically conservative subjects were less likely to punish. Gender and initial contribution have no significant coefficients.[38]

Endogeneity and the effectiveness of IS and non-deterrent FS

Two major departures from the predictions of standard theory in our data are that informal sanctions are effective in deterring free-riding and that non-deterrent formal sanctions are somewhat effective despite being monetarily insufficient to render free-riding unprofitable. While informal sanctions have also been shown to increase contributions in other VCM-with-punishment experiments, their efficiency is atypically high in our data, and not just contributions but also earnings are significantly higher in IS than NS condition for experienced subjects (i.e., in Phase 5).[39]

It seemed possible to us that the high efficiency of IS in our experiment was partly due to the fact that it was chosen by the subjects, who were always offered an alternative

---

[38] As mentioned in footnote 32, in Appendix tables A.5 and A.6 we include two tables investigating the determinants of contribution decisions by individuals under the IS, non-deterrent FS, and NS regimes. With regard to individual characteristics, those regressions show that (a) CRT score is a significant positive predictor of contribution under IS, (b) more politically conservative subjects contribute less under NS, and (c) no significant effects of gender are found. Result (b) corroborates a finding in Putterman *et al*. (2010) (see that paper's Appendix Table B.9).

[39] The difference between average earnings under NS and IS is not statistically significant in phase 2, with the exception of the **NC** treatment, where earnings are significantly higher under IS (p = .07 in two-tailed Mann-Whitney test at group level). In Phase 5, earnings are significantly higher under IS than under NS in the **DC** (p=.001) and **NE** (p=.017) treatments and also when all treatments are pooled (p < .001). In no instance does a test for any treatment in any phase find earnings to be significantly *lower* under IS than under NS. While the tests referred to are between-group tests for given phases, within-group tests for cases in which a given group can be observed under both IS and NS give similar results.

rule to use. Subjects in the **DC** treatment who voted to try out the IS scheme for the first time in Phase 4 may have taken the vote outcome as a signal of a desire to cooperate without incurring the 2 point per period cost of deterrent FS. Sutter *et al*. (2010) find evidence of what they call a "democratic participation-rights premium" with respect to the efficacy of reward and punishment schemes in a public goods game.

We test this conjecture using behaviors in the **Exogenous IS** treatment, in which subjects encountered the rules experienced by counterparts in treatment **DC** in the order that was the most common path leading to a trial of IS in that treatment: a sequence in which subjects played under IS for the first time in Phase 4 following NS in phases 1 and 2 and FS in Phase 3. Whereas use of NS in Phase 2, FS in Phase 3, and IS in Phase 4 was the result of voting for the four groups exhibiting that order in the DC treatment, in **Exogenous IS** the same rules were assigned exogenously, and no mention was made of voting.[40] Notice that the groups to be compared in both the **DC** and **Exogenous IS** treatments used formal deterrent sanctions in Phase 3. If it is experience using a formal sanctions regime which punishes free riding that leads subjects to use informal sanctions more efficiently, and not voting choice, there should be no significant difference in contributions and earnings in Phase 4 for the two sets of groups.

We find the average contribution under IS in Phase 4 of the exogenous treatment to be 14.8 in the 6 groups included, versus 18.8 in Phase 4 for the 4 groups that experienced the same order in the **DC** treatment. Average earnings were also lower in the groups using IS exogenously than in those using it endogenously, at 22.7 versus 35.6. Both differences are

[40] To make the comparison as close as possible, all aspects of the instructions and procedure in the new treatment were identical to those in **DC**, including that subjects were informed of the rules of the informal and formal sanctions schemes after Phase 1 and prior to Phase 2, except that subjects were told that the computer would decide which rule they would be assigned, and were not told on what that decision would be based. While tests using paths observed in other treatments are not ruled out, having limited resources we chose the most commonly observed path leading to an IS condition in **DC** because it includes the largest number of uniform observations on a single path and thus allows for the most high-powered test of its kind. Our test focuses on Phase 4 because thereafter the four DC treatment groups diverge in their voting choices.

significant in two-tailed Mann-Whitney tests, with *p*-values of .019. The conjecture that endogenous (voting) choice contributes to the effectiveness of IS is thus supported.[41]

Efficacy of endogenously chosen non-deterrent sanctions has been documented in the literature, to our knowledge, only by Tyran and Feld (2006), more recently joined by Kamei (2010). It seemed possible to us that, as in the latter experiments, non-deterrent sanctions have an effect in our experiment because their choice by a majority of a group's members signals to each an intention to cooperate. In particular, voting to implement a non-deterrent sanction at a positive cost even though that scheme leaves contributing to the public good a privately unprofitable action might be interpreted by subjects as a signal of intentions to contribute once the group lowers the private cost of that act by imposing the sanction scheme. To test this conjecture, we use the behaviors in the **Exogenous Non-deterrent FS** treatment in which subjects experienced such a sanctions rule and the no sanctions regime in the most common order seen in the **NC** treatment, that in which groups use no sanctions in Phases 1 and 2 and non-deterrent, cheap formal sanctions in Phase 3.[42] As with the test for exogenous versus endogenous IS, the results are supportive of our conjecture, albeit at a lower level of significance: the 9 groups using non-deterrent sanctions exogenously had average contributions of 11.5 points in Phase 3 whereas the 7

---

[41] A possible concern is that subjects who vote for IS differ from those voting for FS and that IS thus performs better in the endogenous than in the exogenous treatment because more of those observed using it in the former are of this IS-preferring type, whereas there is no selection for type among IS users in the exogenous treatment. One way to test for such a selection effect is to check whether pro-IS voters differ from pro-FS voters in their contribution behaviors when IS has been selected for Phase 4 by groups in the **DC** treatment. We find that Phase 4 contributions average 19.2 for pro-IS voters versus 18.6 for pro-FS voters, the difference being insignificant according to a Mann-Whitney test. We can also check whether subjects in the **DC** and exogenous IS treatments, or **DC** subjects voting for IS versus FS in Phase 4, differ in terms of their period 1 contributions, their average contributions in Phase 1, or their gender, CRT score, or political preference. Neither comparison shows a statistically significant difference between groups of subjects with respect to any of these five criteria. The evidence provides reassurance that the observed endogeneity effect is not an artifact of subject selection.

[42] Again, we focus on the single most common path because it allows for the highest-powered test feasible. Basic design features parallel those of **Exogenous IS**, e.g. subjects learn the rules of non-deterrent FS and of IS after Phase 1 play under NS, hear nothing about voting, and are told that whether they play a given phase under NS, FS or IS will be determined by the computer. As with the previous test, we focus on the first phase in which the scheme (in this case, non-deterrent FS) is used—here, Phase 3—because the seven **NC** treatment groups diverge in their choices in Phase 4.

groups using those sanctions endogenously following the NS, NS, IS pathway in **NC** had average contributions of 13.7 points, and the difference is significant with a *p*-value of .064 in a 2-tailed Mann-Whitney test.[43] Since earnings are perfectly correlated with contributions in FS, the difference in average earnings is significant at the same level. Contributions in the 5 groups that chose to use NS rather than FS in Phase 3 of the **NC** treatment average 9.9 points and do not significantly differ from those in the 9 groups experiencing FS exogenously (i.e., those whose average contribution is 11.5) according to a group level Mann-Whitney test (*p*-value = .160), which suggests that imposition (as opposed to voted choice) of non-deterrent formal sanctions fails to increase contributions relative to a condition without sanctions. This second comparison is less ideal than is the test concerning endogenous versus exogenous FS, since in this case it would be better to use observations for groups also encountering NS exogenously.

5. Discussion and Conclusion

Centralization of the power to punish is a hallmark of civilization that, when effective, creates beneficial expectations of relatively certain and well-targeted formal sanctions for violating cooperative norms. However, the threat of horizontal or informal sanctions such as shunning, bad-mouthing, and ostracism remain important spurs to cooperation in many settings. There are numerous collective action problems in which the reach of the state is infeasible or too costly, and the groups concerned may consider using either

---

[43] As in the corresponding case for IS, a possible concern is that subjects using FS endogenously are a self-selected group, since we only observe that condition among groups having a majority of voters favoring FS, whereas subjects using FS exogenously have not been selected for any particular predisposition. Evidence that this is so could take the form that subjects who voted for FS in Phase 3 of the **NC** treatment responded more positively to non-deterrent FS than did ones who voted for NS. What the data show is that subjects who voted for FS contribute an average of 14.0 points to the group account while subjects who voted for NS contribute an average of 14.5, the difference being statistically insignificant according to a Mann-Whitney test. We can also check whether subjects using FS in **NC** treatment are by chance different from those in the exogenous FS treatment in terms of the same characteristics mentioned in footnote 41: period 1 and Phase 1 contributions, gender, CRT score, and political preference. A check for differences on these measures between pro-FS and pro-NS voters in Phase 3 of **NC** treatment finds no significant difference on any of the five. The evidence thus again provides reassurance that the observed endogeneity effect is not an artifact of subject selection.

organizationally administered formal sanctions, informal sanctions, or a combination of the two.[44] Informal sanctions have been extensively studied in the literature, perhaps in response to the surprising discovery that they often increase cooperation when they should not, according to standard economic theory. The ability of informal sanctions to increase social welfare is debated, with some evidence suggesting that they are too costly a remedy to improve efficiency. In contrast, formal sanctions have received almost no attention in the literature, perhaps because it was assumed that they would be as effective as standard economic theory predicts. The comparative performance of formal and informal sanctions, and their respective popular support has not been studied previously, perhaps because it was assumed that formal sanctions would dominate on both accounts. This paper shows that many of these presumptions are overly simplistic, if not wrong. We study the choice between informal and formal sanctioning systems as well as the possibility of confronting the collective action problem without employing sanctions, using four treatments that vary the cost (cheap or expensive) and strength (deterrent or non-deterrent) of formal sanctions.

In line with standard theory, we find that deterrent formal sanctions increase cooperation. In contrast to standard theory (but in line with some recent experimental evidence), we find that informal sanctions also induce high levels of cooperation. In addition, we find that informal sanctions increase efficiency because such sanctions are used diligently and are mostly well-targeted at free riders, in particular with some experience. Because they are used diligently, informal sanctions outperform both costly and cheap formal deterrent sanctions, with experience. Voters tend to anticipate and learn about the high relative cost-effectiveness of informal sanctions, and increasingly vote for informal sanctions over formal ones. The competitiveness of informal with formal sanctions in our experiment is arguably one of the most striking manifestations of the presence and power of the willingness to punish free riders in the informal sanctions literature to date.

---

[44] It may be difficult for any authority to fully rule out informal sanctions, and for that reason among others, informal and formal sanctions may often operate in tandem. For an experimental study of the joint action of the two, see Kube and Traxler (2010). An experimental design like our own amended so that IS opportunities are simultaneously available when a group votes for deterrent or non-deterrent FS would be an interesting extension.

In contrast to the predictions of standard theory, we demonstrate that deterrent formal sanctions may not necessarily dominate a no-sanctions regime. Formal sanctions are worthwhile if the social gain from deterrence exceeds its cost, and we study cases where formal sanctions are expensive or cheap. The fixed cost in our treatments is such that it consumes 40 percent or 10 percent of the potential social gain, defined in standard rather than behavioral terms. However, actual behavior does not coincide with predicted equilibria, with theoretically deterrent sanctions not fully deterrent, and, more importantly, some cooperation even absent any sanction. Both of these factors reduce the actual gain from enacting formal sanctions compared to the potential gain. The parameters we choose for the fixed costs span the tipping point which makes deterrent sanctions behaviorally counterproductive even when they should be efficiency-increasing according to standard economic theory. A majority of subjects appear to realize this and vote accordingly.

The significantly greater effectiveness of informal sanctions when selected by voting compared with their performance in our exogenous comparison treatment gives rise to a second major finding, which is that deciding whether or not to use this regime by majority vote significantly improves its functioning. This result is in line with others regarding the role of coordination in using punishment effectively, as well as with Dal Bó *et al.*'s (2010) finding about the benefits of democratic choice of institutions.[45]

A third finding, paralleling the second, is that costly formal sanctions of non-deterrent magnitude, which should in theory leave the level of cooperation unchanged even as it cuts into profits, can enhance cooperation when selected democratically. Comparison with a treatment in which the same non-deterrent sanctions regime was imposed exogenously shows its efficacy to be significantly higher, reinforcing the result in Tyran and Feld (2006). Indeed, contributions with exogenous non-deterrent sanctions are not significantly higher than those with no sanctions. Whereas informal sanctions prove as effective as deterrent formal sanctions in treatments **DC** and **DE**, however, non-deterrent formal sanctions show less power to induce cooperation than do informal ones (which, in our data, are sufficient in targeting and quantity to render contributing privately optimal).

[45] See also Kamei (2010).

Lest we leap too quickly to generalizing the conclusion that informal sanctions are unambiguously preferable to all but the cheapest deterrent formal sanctions, it is worth bearing in mind that our design limits the degree to which problems of retaliation can arise in the informal sanctions environment. In particular, counter-punishment is probably lessened in our design because there is a single punishment stage in each period and individual group members' identities cannot be tracked from one period to the next. The results in Nikiforakis (2008) and Denant-Boemont *et al*. (2007) suggest that an informal sanctions regime with identifiable punishers and with counter-punishment opportunities might have a considerably more difficult time competing for votes with NS and FS alternatives. Still, our IS regime permits retaliation in the form of perverse or anti-social punishment committed as "blind revenge," and we find evidence that such retaliation was indeed practiced. The competitiveness of IS with costly FS might well hold, although possibly over a narrower domain, in future experiments that makes targeted revenge a more accessible option.

At a higher level of abstraction, informal and formal sanctions should probably not be viewed entirely as alternatives. Rather, the centralized organizational or state structures that make formal sanctions a possibility may require the earlier and perhaps ongoing solution of a prior social dilemma, as seems especially obvious when speaking of a democratic state. At this level, our finding that informal cooperation is surprisingly successful should not be read as favoring informal over formal sanctions in any particular setting, but should be understood, rather, as a testament to the potential for cooperating to create and sustain the state machineries with which formal sanctions become an option in the first place.

Figure 1 Timing in the experiment

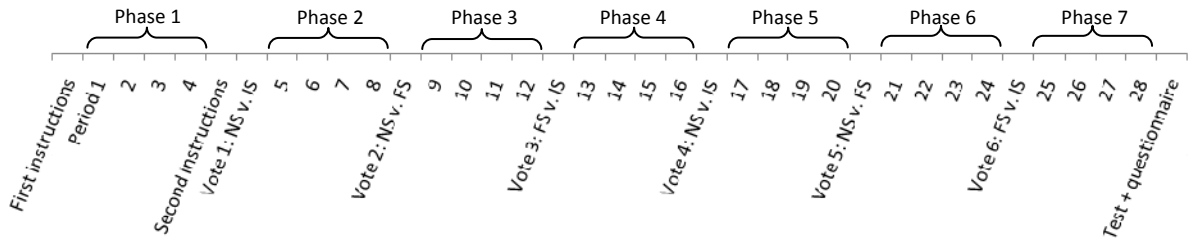
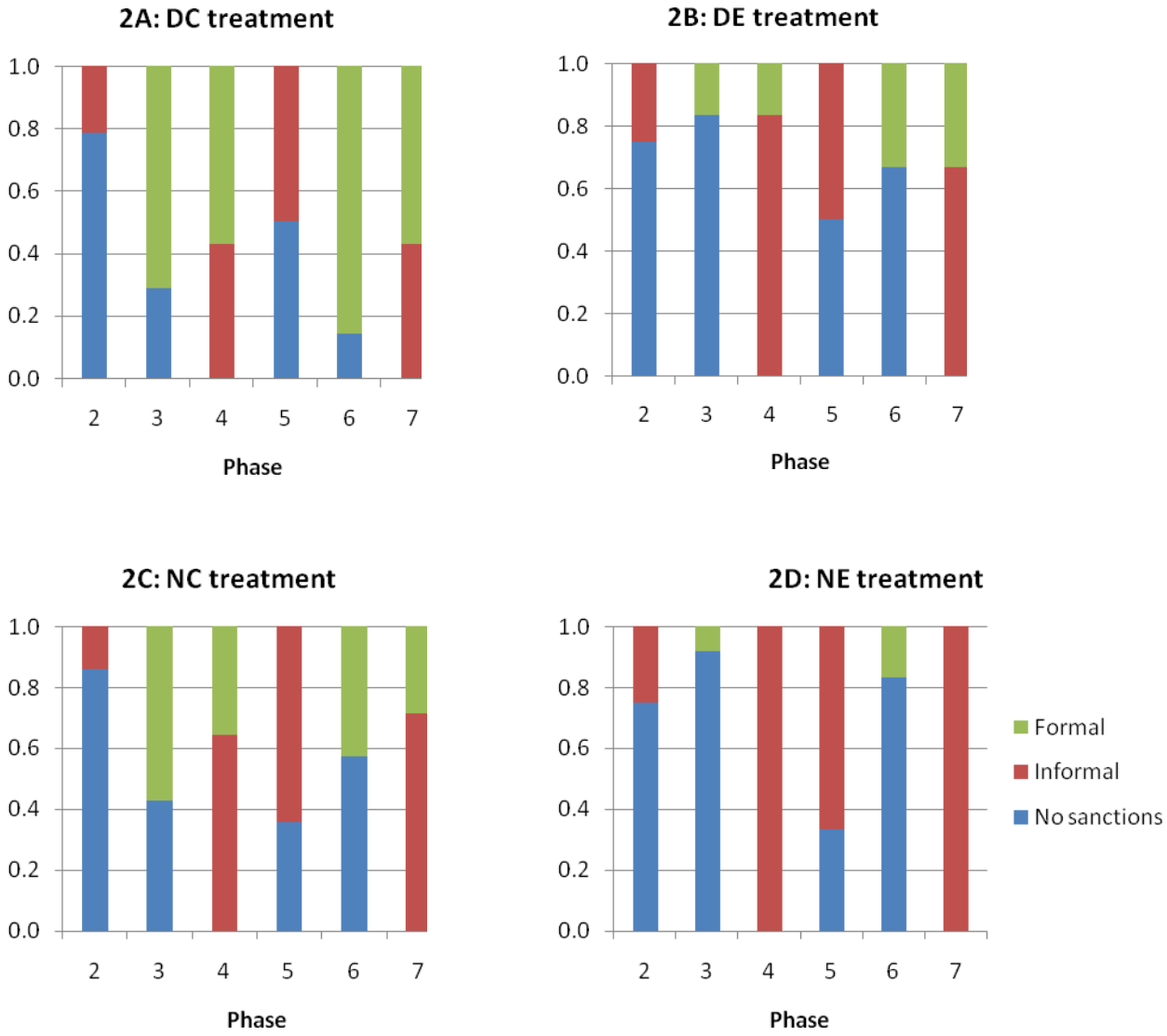


Figure 2 Voting outcomes at the group level



Note: Bars show, for each treatment , the share of *groups* implementing formal, informal and no sanctions in each phase.

Figure 3 Contributions

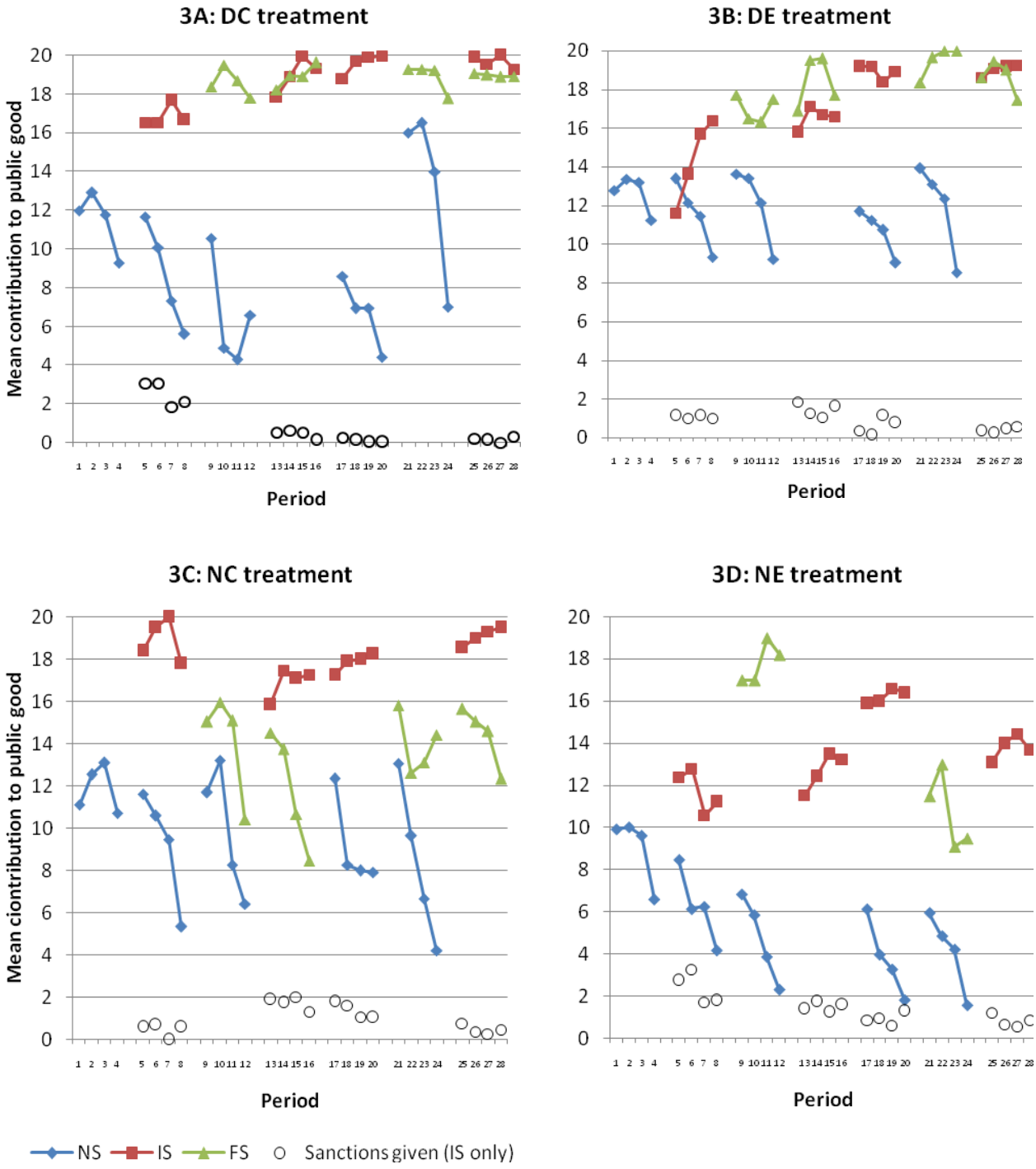
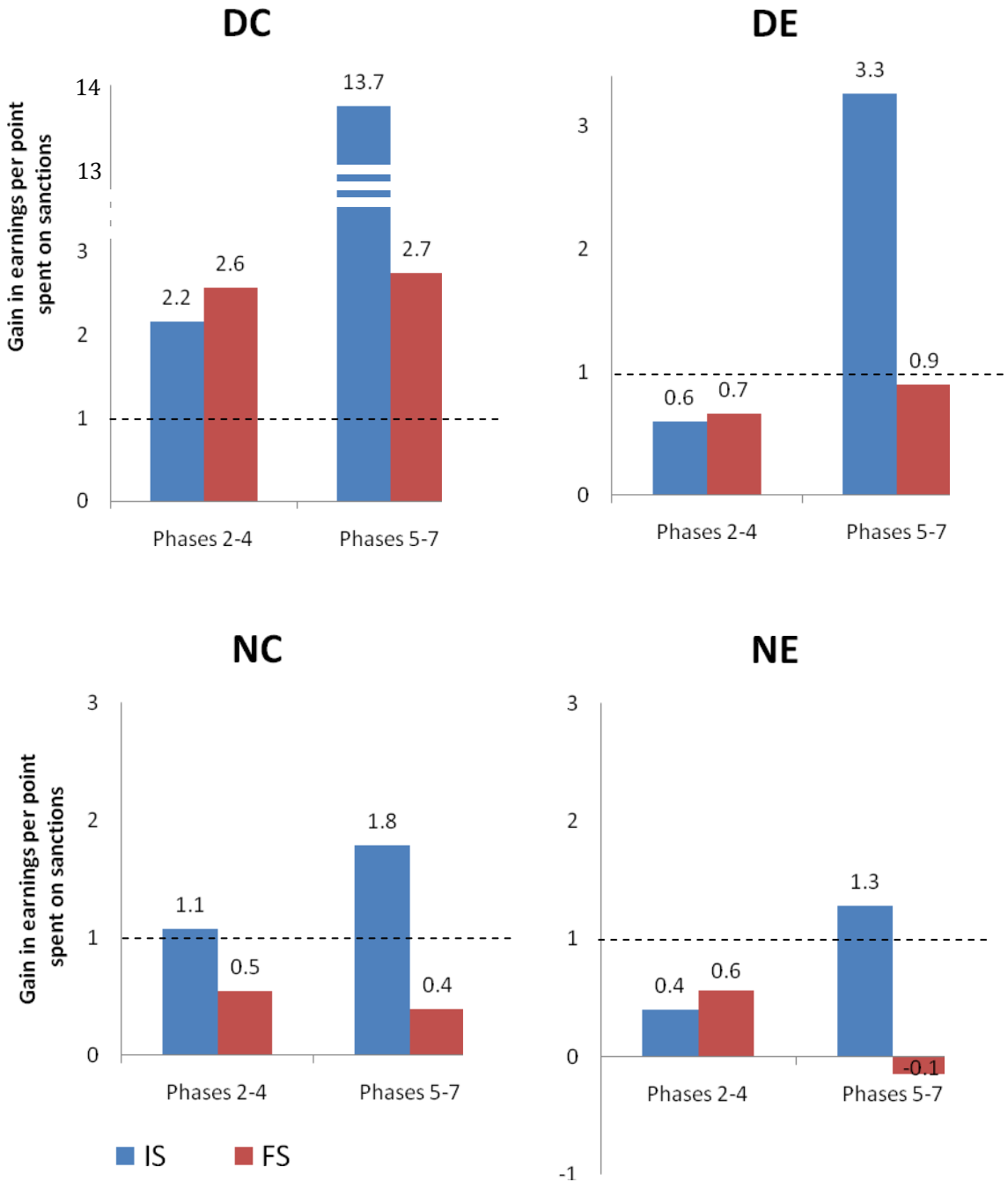


Figure 4 Cost effectiveness of sanctions



Note: Bars show the average gain in gross earnings (i.e. before sanction costs are deducted), relative to earnings in Phase 0, for each point spent on sanctions, including losses to those punished. Earnings in Phase 0 are calculated separately for the groups that experienced the relevant institutions in the relevant phases. Dashed line at gain level 1.0 indicates the break-even point, at which earnings gain equals sanction cost.

Table 1 Treatments (Formal Sanctions Parameters)

	c = 2 "Cheap"	c = 8 "Expensive"	Exogenous Treatments*
s = 0.8 "Deterrent"	Deterrent and Cheap (**DC**)	Deterrent and Expensive (**DE**)	**Exogenous IS** (Informal Sanctions)
s = 0.4 "Non-deterrent"	Non-deterrent and Cheap (**NC**)	Non-deterrent and Expensive (**NE**)	**Exogenous Non-deterrent FS** (Formal Sanctions)

* In both exogenous treatments, c = 2 while the value of s is as indicated by the relevant row heading.

Table 2: Number of subjects and group, by treatment

	c = 2	c = 8	Exogenous Treatments
s = 0.8	(**DC**) Groups: 14 Subjects: 70	(**DE**) Groups: 12 Subjects: 60	**Exogenous FS** Groups: 6 Subjects: 30
s = 0.4	(**NC**) Groups: 14 Subjects: 70	(**NE**) Groups: 12 Subjects: 60	**Exogenous Non-deterrent FS** Groups: 9 Subjects: 45

Note: Each treatment except Exogenous IS had 3 sessions. The main treatments included 52 groups with a total of 260 subjects, with an additional 15 groups having 75 subjects participated in the two exogenous treatments.

Table 3 Effects of cost and deterrence level on voting for formal sanctions

	Dependent variable: Adopted formal sanctions in..							
	Phase 3 (vote 2)		Phase 4 (vote 3)		Phase 6 (vote 5)		Phase 7 (vote 6)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Deterrent FS	0.148	0.158	0.223*	0.204*	0.355**	0.269*	0.334***	0.319**
	(0.146)	(0.147)	(0.122)	(0.123)	(0.139)	(0.160)	(0.121)	(0.152)
Cheap FS	0.525***	0.524***	0.396***	0.343***	0.429***	0.225	0.292**	0.411**
	(0.113)	(0.114)	(0.109)	(0.128)	(0.130)	(0.191)	(0.120)	(0.163)
IS in Phase 2		-0.173		0.033		0.282		-0.050
		(0.168)		(0.163)		(0.193)		(0.188)
FS in Phase 3				0.127		0.392**		-0.195
				(0.144)		(0.185)		(0.171)
FS in Phase 4						0.252		0.200
						(0.202)		(0.210)
IS in Phase 5						-0.313*		-0.579***
						(0.176)		(0.156)
FS in Phase 6								0.046
								(0.189)
Observations	52	52	52	52	52	52	52	52

Probit regressions, marginal effects reported. Units of analysis are groups. Standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 4 Determinants of punishment

	Dependent variable: Amount of punishment given to individual other					
	RE-GLS	RE-GLS	RE-GLS	RE-TOBIT	RE-TOBIT	RE-TOBIT
Others' mean contribution in previous period	0.006	0.007	0.007	-0.054***	-0.042**	-0.041**
	(0.007)	(0.007)	(0.007)	(0.019)	(0.019)	(0.019)
Positive deviation from other's contrib.	0.104***	0.104***	0.082***	0.331***	0.338***	0.325***
	(0.007)	(0.007)	(0.015)	(0.014)	(0.014)	(0.026)
Positive deviation*contribution in period 1			0.002			0.001
			(0.001)			(0.002)
Negative deviation from other's contrib.	0.039***	0.038***	0.046	0.131***	0.124***	0.115***
	(0.013)	(0.013)	(0.030)	(0.014)	(0.014)	(0.023)
Negative deviation*contribution in period 1			-0.001			0.001
			(0.002)			(0.002)
Non-perversely punished in previous period	0.035	0.027	0.030	0.458***	0.362***	0.366***
	(0.025)	(0.025)	(0.026)	(0.127)	(0.127)	(0.127)
Perversely punished in previous period	-0.105	-0.103	-0.097	-0.326	-0.308	-0.305
	(0.067)	(0.066)	(0.061)	(0.243)	(0.244)	(0.244)
CRT score		-0.089***	-0.089***		-0.578***	-0.576***
		(0.026)	(0.025)		(0.085)	(0.085)
Female		0.007	0.013		-0.140	-0.144
		(0.072)	(0.069)		(0.172)	(0.172)
General political preference		-0.010	-0.010		-0.136***	-0.137***
		(0.014)	(0.013)		(0.034)	(0.033)
Contribution in period 1		-0.001	-0.002		-0.012	-0.017
		(0.005)	(0.003)		(0.012)	(0.014)
Phase:						
Phase 4	-0.028	-0.035	-0.034	-0.192	-0.268	-0.269
	(0.033)	(0.031)	(0.032)	(0.188)	(0.188)	(0.187)
Phase 5	-0.024	-0.029	-0.031	-0.331	-0.377*	-0.377*
	(0.039)	(0.039)	(0.040)	(0.206)	(0.206)	(0.206)
Phase 7	-0.099*	-0.106**	-0.104*	-1.004***	-1.070***	-1.071***
	(0.054)	(0.053)	(0.054)	(0.208)	(0.208)	(0.208)
Period:						
Period 3	-0.016	-0.017	-0.020	-0.188	-0.208*	-0.209*
	(0.025)	(0.025)	(0.025)	(0.120)	(0.120)	(0.120)
Period 4	-0.009	-0.01	-0.011	-0.201*	-0.220*	-0.223*
	(0.038)	(0.038)	(0.038)	(0.120)	(0.121)	(0.121)
Treatment:						
DC	-0.072	-0.067	-0.061	-0.880***	-0.861***	-0.875***
	(0.080)	(0.083)	(0.083)	(0.276)	(0.270)	(0.271)
DE	-0.017	-0.006	-0.010	-0.484**	-0.375	-0.380
	(0.087)	(0.084)	(0.083)	(0.241)	(0.236)	(0.236)
NC	-0.003	0.005	0.000	-0.283	-0.266	-0.275
	(0.090)	(0.093)	(0.094)	(0.235)	(0.230)	(0.231)
Constant	0.078	0.252	0.263*	-2.295***	-0.696	-0.651
	(0.105)	(0.154)	(0.148)	(0.353)	(0.431)	(0.435)
R-sq- (overall)	0.21	0.21	0.22			
Log-likelihood				-3149.82	-3120.78	-3120.54
Observations	9120	6840	6840	6840	6840	6840
Number of dyads	940	940	940	940	940	940

Note: Standard errors in parentheses. Standard errors corrected for within-group clustering in GLS- but not in tobit regressions. CRT score, gender, political orientation and period 1 contribution refer to *senders* of punishment, not to receivers. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 5 Voting and earnings

	Dependent variable:						
	Voted for *formal* sanctions (against informal) in **Phase 4**	Voted for *informal* sanctions (against no sanctions) in **Phase 5**		Voted for *formal* sanctions (against no sanctions) in **Phase 6**		Voted for *formal* sanctions (against informal) in **Phase 7**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Earn_fs/earn_is	0.709**					0.849***	0.677***
	(0.322)					(0.254)	(0.228)
Earn_is/earn_ns		0.865***	0.889***				
		(0.132)	(0.148)				
Earn_fs/earn_ns				0.566***	0.580***		
				(0.202)	(0.207)		
CoV_is/CoV_ns			0.002				
			(0.003)				
CoV_fs/CoV_ns					0.015		
					(0.035)		
CoV_fs/CoV_is							-0.043
							(0.059)
Treatment:							
DC		0.038	0.039	0.136	0.144	0.032	0.183
		(0.112)	(0.114)	(0.121)	(0.123)	(0.172)	(0.188)
DE		-0.079	-0.031	0.100	0.110	-0.173	0.113
		(0.097)	(0.095)	(0.157)	(0.157)	(0.177)	(0.219)
NC		-0.114	-0.111			-0.126	-0.08
		(0.110)	(0.111)			(0.145)	(0.162)
Pseudo-Rsq	0.4	0.21	0.22	0.13	0.13	0.42	0.43
Observations	15	200	195	125	125	125	85

Note: Probit regressions, marginal effects reported. Standard errors in parentheses. Standard errors adjusted for within-group clustering, except in the regression for Phase 4, where only three group are included. In the regression for Phase 6, the outcome does not vary within the **NE** treatment (all five individuals with the relevant experience voted for formal sanctions) and the observations in this treatment can therefore not be included. The earnings variables are always based on experience from *the most recent phase* where the subject experienced the institution in question. For example, "earn_fs/earn_is" is mean earnings in the most recent phase with formal sanctions, divided by mean earnings in the most recent phase with informal sanctions. "CoV" stands for "Coefficient of Variation". The number of observations in column (7) is less than that in column (6) due to the dropping of observations for which CoV_is = 0. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 6 Earnings by treatment, condition and phase

		Phase						
Treatment	*Condition*	1	2	3	4	5	6	7
DC	NS	126.0	114.7	106.4		106.9	133.5	
	IS		96.7		146.6	155.1		154.9
	FS			141.8	144.4		144.0	144.6
	*MW-test, p-value**		*0.39*	*0.01*	*0.37*	*0.00*	*0.14*	*0.02*
	% choosing higher earning option		78.6%	71.4%	42.9%	50.0%	85.7%	42.9%
DE	NS	130.6	126.3	128.4		122.7	127.9	
	IS		116.0		117.8	144.1		148.2
	FS			106.4	116.7		124.4	118.1
	MW-test, p-value		*0.93*	*0.28*	*0.28*	*0.15*	*0.73*	*0.02*
	% choosing higher earning option		75.0%	83.3%	83.3%	50.0%	66.7%	66.7%
NC	NS	127.5	117.1	119.6		116.6	113.6	
	IS		146.2		113.1	125.8		147.1
	FS			119.2	106.4		118.4	120.7
	MW-test, p-value		*0.07*	*0.70*	*0.73*	*0.46*	*0.30*	*0.02*
	% choosing higher earning option		14.3%	42.9%	64.3%	64.3%	42.9%	71.4%
NE	NS	116.2	105.1	99.0		95.3	96.7	
	IS		78.9		99.7	125.7		118.5
	FS			115.7			76.3	
	MW-test, p-value		*0.23*	*0.11*	-	*0.02*	*0.03*	-
	% choosing higher earning option		75.0%	8.3%	0.0%	66.7%	83.3%	100.0%
All	NS	125.2	115.8	113.4		111.3	113.1	
	IS		106.1		115.5	136.3		139.1
	FS			128.6	128.0		128.7	132.0
	MW-test, p-value		*0.7*	*0.0*	*0.4*	*0.0*	*0.0*	*0.1*
	% choosing higher earning option		59.6%	51.9%	48.1%	57.7%	69.2%	69.2%

*Mann-Whitney tests of the hypothesis that earnings under the two conditions are equal. Tests are conducted at the *group* level, for phase averages. All Mann-Whitney tests reported here and elsewhere in the paper are two-tailed.

References

Andreoni, James, 1988, "Why Free Ride? Strategies and Learning in Public Goods Experiments," *Journal of Public Economics* 37: 291-304.

Andreoni, James, 1993, "An Experimental Test of the Public Goods Crowding-Out Hypothesis," *American Economic Review* 83: 1317-1327.

Bochet, Olivier, Talbot Page and Louis Putterman, 2006, "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior and Organization*, 60: 11-26.

Botelho, Anabela, Glenn Harrison, Ligia M. Costa Pinto and Elisabet E. Rutström, 2005, "Social Norms and Social Choice," unpublished paper, Dept. of Economics, University of Central Florida.

Boyd, Robert, Herbert Gintis and Samuel Bowles, "Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate when Rare," 2010, *Science* 328: 617-620 (March-April).

Buchanan, James and Gordon Tullock, 1962, *The Calculus of Consent. Logical Foundations of Constitutional Democracy.* Ann Arbor: University of Michigan Press.

Chaudhuri, Ananish, forthcoming, "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature," *Experimental Economics* (in press).

Cinyabugama, Matthias, Talbot Page and Louis Putterman, 2006, "Can Second-Order Punishment Deter Perverse Punishment?" *Experimental Economics* 9: 265-79.

Chaney, William, 1963, "Anglo-Saxon Church Dues: A Study in Historical Continuity," *Church History* 32 (3): 268-277.

Dal Bó, Pedro, Andrew Foster and Louis Putterman, 2010, "Institutions and Behavior: Experimental Evidence on the Effects of Democracy," *American Economic Review* (Dec., in press).

Davis and Holt, 1993, *Experimental Economics*. Princeton: Princeton University Press.

Denant-Boemont, Laurent, David Masclet and Charles Noussair, 2007, "Punishment, Counter-punishment and Sanction Enforcement in a Social Dilemma Experiment," *Economic Theory* 33: 145-167.

Egas, Martijn and Arno Riedl, 2008, "The Economics of Altruistic Punishment and the Maintenance of Cooperation," *Proceedings of the Royal Society B* 275: 871-878.

Ertan, Arhan, Talbot Page and Louis Putterman, 2009, "Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free-Rider Problem" *European Economic Review* 53: 495-511, 2009.

Fehr, Ernst and Simon Gächter, 2000a, "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives* 14 (3): 159-81.

Fehr, Ernst and Simon Gächter, 2000b, "Cooperation and Punishment in Public Goods Experiments," *American Economic Review* 90: 980-994.

Fehr, Ernst and Simon Gächter, 2002, "Altruistic Punishment in Humans," *Nature* 415: 137-140.

Fehr, Ernst and Klaus Schmidt, 1999, "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics* 114: 817-68.

Fischbacher, Urs, 2007, "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics* 10: 171-178.

Fischbacher, Urs and Simon Gächter, 2010, "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments," *American Economic Review*, forthcoming.

Frederick, Shane, 2005, "Cognitive Reflection and Decision-Making," *Journal of Economic Perspectives* 19: 25-42.

Gächter, Simon and Benedikt Herrmann, 2009, "Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment," *Philosophical Transactions of the Royal Society B* 364: 791-806.

Gürerk, Ő., B. Irlenbusch and B. Rockenbach, 2006, "The Competitive Advantage of Sanctioning Institutions," *Science* 312 pp. 108-110, April 7 2006.

Herrmann, Benedikt, Christian Thöni and Simon Gächter, 2008, "Antisocial Punishment Across Societies," *Science* 319: 1362-7.

Hichri, W., 2004, "Interior Collective Optimum in a Voluntary Contribution to a Public-Goods Game," *Applied Economics Letters* 11: 135-140.

Hobbes, Thomas, 1996 [1651], *Leviathan. Or the Matter, Forme and Power of a Commonwealth Ecclesiastical and Civil*. New York: Oxford University Press.

Janssen, Marco, Robert Holahan, Allen Lee and Elinor Ostrom, 2010, "Lab Experiments for the Study of Social-Ecological Systems," *Science* 328: 27-36 (March-April).

Kamei, Kenju, 2010, "Democracy and Resilient Pro-Social Behavioral Change: An Experimental Study," Brown University Department of Economics (available at www.econ.brown.edu/students/kenju_kamei/JMP.pdf).

Kamei, Kenju, Louis Putterman and Jean-Robert Tyran, 2011, "State or Nature? Formal vs. Informal Sanctioning in the Voluntary Provision of Public Goods," Brown University Department of Economic Working Paper No. 2011-3.

Kreps, David, Paul Milgrom, John Roberts and Robert Wilson, 1982, "Rational Cooperation in Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory* 27: 245-52.

Kube, Sebastian and Christian Traxler, 2010, "The Interaction of Legal and Social Norm Enforcement," CESifo Working Paper No. 3091.

Ledyard, John O., 1995, "Public Goods: A Survey of Experimental Research." in John Kagel and Alvin Roth, eds., *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

Locke, John, 2005 [1689], *Two Treatises of Government and a Letter Concerning Toleration.* Digireads.com Publishing, Stilwell.

Nikiforakis, Nikos, 2008, "Punishment and Counter-punishment in Public Good Games: Can we Really Govern Ourselves?" *Journal of Public Economics* 92: 91 – 112.

Nikiforakis, Nikos and Hans Normann, 2008, "A Comparative Statics Analysis of Punishment in Public Goods Experiments," *Experimental Economics* 11: 358-369.

Önes, Umut and Louis Putterman, 2007, "The Ecology of Collective Action: A Public Goods and Sanctions Experiment with Controlled Group Formation," *Journal of Economic Behavior and Organization* 62: 495-521.

Ostrom, Elinor, 2010, "Beyond Markets and States: Polycentric Governance of Complex Economic Systems," *American Economic Review* 100: 641 – 672.

Ostrom, Elinor, James Walker and Roy Gardner, 1992, "Covenants with and without a Sword: Self Governance is Possible." *American Political Science Review.* 86 (2): 404-416.

Palfrey, Thomas and Prisbrey, Jeffrey, 1997, "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *American Economic Review*; 87(5): 829-46.

Putterman, Louis, Jean-Robert Tyran and Kenju Kamei, 2010, "Public Goods and Voting on Formal Sanction Schemes: An Experiment," Working Paper, Department of Economics, Brown University and University of Copenhagen.

Reuben, Ernesto and Arno Riedl, 2009, "Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations," Discussion Paper, University of Maastricht.

Rockenbach, Bettina and Irenaeus Wolff, 2009, "Institution Design in Social Dilemmas: How to Design if You Must?" Working Paper, University of Erfurt.

Sefton, Martin, Robert Shupp and James Walker, 2007, "The Effect of Rewards and Sanctions in Provision of Public Goods," *Economic Inquiry* 45: 671-690.

Selten, Reinhard, 1975, "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 4: 25 – 55.

Sen, Amartya, 1967, "Isolation, Assurance and the Social Rate of Discount," *Quarterly Journal of Economics* 81: 112-124.

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

Thöni, Christian, Jean-Robert Tyran and Erik Wengström, 2009, "Microfoundations of Social Capital," Working Paper 09-24, Department of Economics, University of Copenhagen.

Tyran, Jean-Robert and Lars P. Feld, 2006, "Achieving Compliance when Legal Sanctions are Non-deterrent," *Scandinavian Journal of Economics* 108 (1): 1-22.

Zelmer, Jennifer, 2003, "Linear Public Goods Experiments: A Meta-Analysis," *Experimental Economics* 6: 299-310.