Self-Organization for Collective Action

An Experimental Study of Voting on Sanction Regimes[*]

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, yet informal or horizontal sanctions have attracted more attention of late. We study experimentally a collective action dilemma and test whether subjects choose a formal sanction scheme that costs less than the surplus it makes possible, as predicted by standard economic theory, or instead opt for the use of informal sanctions or no sanctions. Our subjects choose, and succeed in using, informal sanctions surprisingly often, their voting decisions being responsive to the cost of formal sanctions. Adoption by voting enhances the efficiency of both informal sanctions and non-deterrent formal sanctions. Results are qualitatively confirmed under several permutations of the experimental design.

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

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

[*] We wish to thank Florian Spitzer and Ivana Grujic for excellent research assistance. J.-R. Tyran thanks the Danish research council (FSE) and the Austrian Science Fund (FWF project no. S10307-G16) for financial support. We thank seminar participants at the Universities of Cologne, Copenhagen, Hamburg, Heidelberg, Lausanne, Munich and Oslo for helpful comments. We are grateful to Kenju Kamei for stimulating discussions and thank the managing editor and three anonymous referees for helping us to improve the paper.

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

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

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

1. INTRODUCTION

Classic social thought emphasizes two responses to the possible tensions between the collective interests of groups or societies and the private interests of individuals or families. First, it argues that in many market interactions, the pursuit of private interests serves the general good without actors' conscious intent (Smith, 2003 (1776)). Second, it concludes that where private and collective interests clash—the domain of "social dilemmas"—the greater good is served by ceding some authority to the state which, for example, can fund public goods provision by tax collection (Hobbes 1996 (1651), Locke 2005 (1689)). Economists also devote attention to social dilemmas that groups navigate without the aid of formal authority, for instance the elicitation of effort in work teams and partnerships, and small-scale collective action in communities and organizations. Still, the state enjoys pride of place, where thinking about social dilemmas is concerned, in both received economic theory and classical social thought.

It seems odd, then, that while dozens of recent studies (beginning with Fehr and Gächter, 2000, and surveyed by Chaudhuri, 2010) have considered the mitigation of voluntary collective action problems by informal sanctioning mechanisms, almost no attention has been paid to the comparison of informal sanctions with the formal sanctions characteristic of the state. John Locke argued that sanctioning should be the job of the state due to the problem of retaliatory punishment:

> … 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.) … resistance many times makes the punishment dangerous, and frequently destructive, to those who attempt it. (§. 126.)

Retaliatory, anti-social, and perverse or non-efficiency-enhancing punishment have indeed been key issues in the study of informal sanctions (Nikiforakis 2008, Denant-Boemont, Masclet and Noussair 2007, Cinyabuguma, Page and Putterman 2006 and Herrmann, Thöni and Gächter 2008). Assigning the right to punish to the state has been deemed a hallmark of civilization.

In this paper, we compare the performance of formal and informal sanctioning and study preferences between them as well as the option of operating in a sanction-free environment. We conduct laboratory experiments in which subjects face a linear, finitely repeated voluntary contribution problem and are given opportunities to vote on pairs of alternative schemes.[1] 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 with and without informal sanctions.

Our four main treatments result from the 2x2 crossing of variation in the costs of formal penalties (a) to those punished and (b) to the group as a whole. (a) In two treatments, the penalty for non-contribution 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" (Tyran and Feld 2006). Non-deterrent sanctions are of interest because, while possibly at odds with the spirit of Hobbes' *Leviathan*, they are arguably more common than deterrent ones in real world settings.[2] (b) In two treatments, use of formal sanctions carries a substantial fixed cost to the group, whereas in two others, it carries a much lower fixed cost. Cost is a key issue since up-front expenditure to establish monitoring and enforcement structures like police and a judicial system is an essential feature distinguishing a formal from an informal sanctions regime. A deterrent formal sanction scheme involving either of the two costs should be rationally favored by individuals, according to standard theory, but this might fail to be the case if those individuals behave more cooperatively than predicted in sanction-free or informal sanction environments (as predicted under some parameter settings of well-known social preference theories, discussed below). Because the efficacy of both informal and non-deterrent formal sanction regimes may depend on the concurrence of those operating under them, our experiment checks for a "democratic dividend" by comparing performance under the endogenously chosen vs. exogenously imposed sanctions regimes.

[1] Voting on informal sanctions has been studied in experiments by Botelho *et al*. (2005), Ertan *et al*. (2009) and Sutter *et al*. (2010), while voting on whether to use formal sanctions has been studied by Tyran and Feld (2006), Kosfeld *et al*. (2009) and Kamei (2011). Kroll *et al*. (2007) study voting on an obligation to contribute when the availability of informal sanctions is exogenously determined. Our experiment is the first in which subjects choose between formal and informal sanctions schemes by voting (see Traulsen et al., 2012 and Zhang et al., 2013 for individual choice by "voting with the feet").

[2] One reason is that the penalty required to achieve deterrence may be considered to exceed 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 violating the rule brings informal as well as formal penalties, e.g. social disapproval.

We find, perhaps unsurprisingly, that a majority of subjects vote for and achieve high efficiency with low-cost and deterrent formal sanctions. We find that most subjects are initially disinclined to allow informal sanctions, 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*. 2006, Ertan *et al*. 2009), that informal sanctions become increasingly popular as experience of the collective action dilemma increases.

More surprisingly, we find that subjects tend to prefer informal over even deterrent formal sanctions when the latter are somewhat costly. For example, between 40 and 70 percent of groups select informal over formal sanctions in the final match-up of the two options. And these choices turn out to be smart. In fact, choosing informal sanctions is profitable because they are used with circumspection and therefore quite effectively deter free-riding. We also find that the popularity of formal sanctions depends more on their fixed costs than on their deterrence. For example, about 70 percent prefer no sanctions over deterrent but costly formal sanctions. These findings are surprising from the perspective of standard theory which assumes that decision makers and voters are rational and strictly self-interested. Yet, these (and some other) observations from our experiment are consistent with theories postulating that decision makers have social preferences.

We use a simple model of aversion to inequality (Fehr and Schmidt 1999) to demonstrate that these findings can be rationalized by assuming social preferences (in the appendix, we show that the model of Charness and Rabin, 2002, also predicts many aspects of our findings rather well). Our findings thus add to the accumulating evidence that social preferences facilitate voluntary collective action and render it considerably more feasible than was supposed by earlier theories.[3] Importantly, our experiment provides novel evidence that voters manage surprisingly well to self-organize for collective action, and we thus provide a remarkable example of efficient endogenous emergence of institutions.

[3] To be sure, Buchanan and Tullock wrote half a century ago 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).

Our paper also adds to a small experimental literature suggesting that there is a "dividend of democracy" in the sense that institutions chosen by vote can be more efficient than when the same institutions are exogenously imposed on decision makers (Tyran and Feld 2006, Dal Bó, Foster and Putterman 2010 and Sutter, Haigner and Kocher 2010). In particular, we find that both informal sanctions and non-deterrent formal sanctions are more efficient when collectively chosen by majority vote than when the same sanctioning institution is imposed on subjects.

The surprising popularity of informal sanctions hinges on their comparative effectiveness, in addition to their natural cost advantage (they do not require the costly infrastructure typical for formal sanctions like police, courts, or prisons). The effectiveness is partly driven by the sanctioning technology (i.e. how much harm a punisher can inflict on the punished at a given cost to himself) and partly by circumspection in usage (i.e. whether costly sanctioning is common and targeted at free riders). In control treatments, we test the robustness of our findings to reducing the effectiveness of informal sanctions by 50 percent., We find that this reduces the efficiency of informal sanctions somewhat. But they are still quite popular, in particular if the alternative is costly deterrent formal sanctions. Further control treatments confirm that our main results are robust to learning, and to allowing for the possibility that formal and informal sanctions co-exist.

The rest of the paper proceeds as follows. Section 2 discusses collective choice between formal and informal sanctions regimes from the perspectives of both standard theory and social preference models. Section 3 describes our experimental design, and Section 4 presents the main results. Section 5 discusses the robustness of our findings. Section 6 provides concluding remarks.

2. TRADITIONAL AND SOCIAL PREFERENCE THEORIES

Consider a group of *n* individuals who obtain utility from consuming a private or a public good. Consuming the private good yields private benefits only, while contributing to the public good benefits all *n* group members. We study a situation in which private and social incentives conflict: Contributing to the public good is socially efficient ($mn > 1$), but individual incentives are stacked against contributing ($m < 1$).[4] Individual monetary payoffs thus induce free-rider incentives and create a social dilemma situation:

$$\pi_i = \left(E - C_i\right) + m\sum_{j=1}^{n} C_j \quad , \tag{1}$$

where *E* 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 $1/n < m < 1$, where *n* is group size, so that the socially optimal payoff per person, *mnE*, exceeds the payoff when each individually optimizes, *E*.

When a formal sanction scheme is in place, an individual incurs a sanction of *s* units for each unit she allocates to private rather than group production. Operating the formal sanction scheme requires payment of a fixed cost of $c < (mnE - E) = \mathscr{P}$ up front. The right hand side of the inequality is the "cooperation premium," that is, the difference between an individual's earnings under full cooperation with zero fixed cost and earnings under individual optimization. When the scheme is adopted, *i*'s payoff becomes

$$\pi_i = \left(E - C_i\right)\left(1 - s\right) + m\sum_{j=1}^{n} C_j - c \quad , \tag{2}$$

where *s* is the sanction per unit allocated to the private good.

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 group production. Accordingly, presence of the penalty changes equilibrium play among rational, self-interested agents with common knowledge of type from $C_i = 0$, all

[4] In a more general model, the agents might have an interior optimum in which some of each good is provided, reflecting the fact that public and private goods are usually not perfect substitutes. The simpler set-up followed here, and in most of the experimental literature, captures the essential issue of free riding incentives while reducing complexity for the decision makers.

i, to $C_i = E$, all *i*. Since $mnE - c > E$ (equivalently $\mathscr{P} > c$), each enjoys a higher payoff with the scheme than without it, and it is therefore a dominant strategy to vote for the scheme if any chance of being a pivotal voter is perceived.

When $0 < s < (1 - m)$, we say (following Tyran and Feld, 2006) we have a **non-deterrent formal sanction** because a rational payoff-maximizing individual selects $C_i = 0$ despite the presence of a sanction. Standard theory predicts that a non-deterrent sanction scheme that adds to cost but fails to change behavior is turned down by a pivotal voter.

With an **informal sanction** scheme, each group member has the opportunity to impose costly punishment on other group members at her discretion after seeing how much they have contributed to the group activity. Specifically, any individual *i* can impose a sanction σ on any other group member *j* at a cost of one unit to herself. We denote *i*‘s cost of punishing *j* by R_{ij} , and *j*‘s cost of being punished by *i*, by σR_{ji} . The payoff of an individual *i* under the informal sanctions scheme therefore is

$$\pi_i = \left(E - C_i\right) + m\sum_{j=1}^{n} C_j - \sum_{j=1}^{n} R_{ij} - \sigma\sum_{j=1}^{n} R_{ji} \quad . \qquad (3)$$

It is easy to see that in a one-time interaction, a rational individual *i* seeking to maximize her payoff will not punish at all, i.e. choose $R_{ij} = 0$, all *j*. 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 that informal sanctions are completely irrelevant (no punishment, no contribution to the group activity ($C_i = 0$, all *i*), and a payoff of *E* for each individual). Rational voters are therefore completely indifferent between no sanctions and informal sanctions. As a result, the probability that an informal sanctions scheme is selected by vote is in theory 0.5.

In summary, standard theory predicts full free riding ($C_i = 0$) in the absence of sanctions, under non-deterrent formal sanctions, and under informal sanctions (which are not meted out if present, $R_{ij} = 0$). Standard theory predicts that voters accept deterrent formal sanctions when $c < \mathscr{P}$, reject non-deterrent formal sanctions at any $c > 0$, and are indifferent between informal and no sanctions. These predictions contrast with a large body of experimental results that shows only partial free riding absent sanctions, widespread use of informal sanctions, a positive response of contributions to both informal sanctions and non-deterrent formal sanctions, and some voting for non-deterrent formal sanctions.[5] Importantly, the standard theory also fails to explain some of the main findings of this paper, including the low popularity of deterrent sanctions at moderate fixed costs and the high popularity of informal sanctions with voters.

2.1 *Predictions from a model of aversion to inequality*

This section argues that the main regularities observed in our experiment can be rationalized by a model of social preferences proposed by Fehr and Schmidt (1999). The model assumes that people are all self-interested but that some are also more or less averse to inequality, usually assigning greater weight to inequalities that disadvantage them. Our intention here is not to argue that this particular model is the most suitable to analyze voting on sanctions or to provide a comparative evaluation of alternative theories (see e.g. Tyran and Sausgruber 2006 for an application of the model to voting on redistribution). We focus on it as an example of a wider class of social preference models because it is parsimonious and provides predictions that are qualitatively well in line with our findings. Below, we only provide a sketch of the argument.[6]

Fehr and Schmidt (1999) assume preferences of the form

$$U_i\ (\pi) = \pi_i - \frac{\alpha_i}{n-1}\sum_{j\neq i}\max\left(\pi_j - \pi_i, 0\right) - \frac{\beta_i}{n-1}\sum_{j\neq i}\max\left(\pi_i - \pi_j, 0\right) \qquad (4)$$

with $\alpha_i \geq \beta_i$ and $0 \leq \beta_i \leq 1$. They show that if aversion to advantageous inequality (β) is sufficiently strong among all group members, positive contributions C are an equilibrium in our setting even in the absence of sanctions. They also show that individuals with sufficiently strong aversion to disadvantageous inequality (α) are willing to punish free riders, and that this threat of punishment can induce positive contributions in the presence of informal sanctions. Given empirically plausible distributions of α and β (see e.g. Blanco *et*

[5] For a survey, see Chaudhuri (2010). Voting on non-deterrent sanctions is studied by Tyran and Feld (2006) and Kamei (2011).

[6] Appendix B provides the details of the argument and compares to predictions with two models suggested by Charness and Rabin (2002), one of which also allows for reciprocity. For the sake of simplicity, the reasoning in this section abstracts from repetition and learning in that we assume a one-shot interaction and that preferences and payoffs are common knowledge.

al. 2011), equilibria with positive contributions are therefore more likely to obtain with than without informal sanctions.

Deterrent sanctions are behaviorally robust in the sense that the predictions are largely independent of whether inequality-aversion is present or not. Under weak assumptions, full contribution by all players is the unique equilibrium in either case (see Appendix B). However, assuming social preferences does make a difference for predicting the effects of non-deterrent formal sanctions. For plausible distributions of α and β, equilibria with positive contributions are more likely in the presence than in the absence of such sanctions. The reason is that sanctions come at a fixed cost and reduce the (monetary) returns from private allocation but leave the (psychological) return from reducing inequality unaffected.[7]

Concerning voting, the Fehr-Schmidt model predicts that the popularity of a formal sanction scheme depends on both the severity of sanctions s and the fixed cost associated with their adoption c. In fact, deterrent formal sanctions may fail to be preferred to a regime with no sanctions despite $c < \mathscr{P}$. The premium $\mathscr{P}$ is large according to standard theory because it is the difference between earnings at (universal) full cooperation and earnings at zero cooperation. In contrast, Fehr and Schmidt predict the "behavioral cooperation premium" with deterrent formal sanctions to be smaller than $\mathscr{P}$ because agents voluntarily cooperate absent sanctions if they are sufficiently averse to advantageous inequality. In the no-sanctions case, inequality-aversion tends to generate multiple equilibria: the desire to avoid deviations from other group members' pay-off implies a preference for contributing the same amount to the public good as others. Therefore, equilibria with high and equilibria with low contributions often co-exist, meaning that groups face a coordination problem. If full contribution is assumed under deterrent formal sanctions, groups using no sanctions must coordinate on contributing at least $C' = E - c/(nm-1)$ to make that institution a profitable alternative. As a result, Fehr and Schmidt predict that multiple, symmetric equilibria exist when choosing between deterrent sanctions and no sanctions, provided that agents are

[7] Non-deterrent formal sanctions are in our design equivalent to an increase in the MPCR, which has been shown to increase contributions (e.g. Isaac and Walker 1988). Since a subject who allocates everything to the private good earns less under non-deterrent formal sanctions than under no sanctions, non-deterrent sanctions decrease the overall stakes of the game. Kocher *et al*. (2008) find no effect of stake size on contributions in public goods games with and without informal sanctions.

sufficiently inequity averse. Note that there is an inverse relation between the threshold *C'* and the fixed cost *c*, i.e. higher fixed costs means that the threshold at which group members are indifferent between formal sanctions and no sanctions is lower. If we add the plausible assumption that a lower threshold is more easily reached, the Fehr-Schmidt model thus predicts that inequality-averse voters are more likely to prefer no sanctions over formal sanctions if the latter come at a high fixed cost.

Under non-deterrent formal sanctions, equilibria with positive contributions are feasible for empirically plausible distributions of α and β, when they are not, in the absence of such sanctions. In such cases, the Fehr-Schmidt model predicts voting for non-deterrent formal sanctions provided that the gains from the cooperation it induces exceed the cost *c*.

Because equilibria with positive contributions are more likely under informal than under no sactions, the potential of costly formal sanctions to justify their fixed cost falls further when the alternative is informal sanctions. The Fehr-Schmidt model therefore predicts fewer votes for formal sanctions when pitted against informal ones than when pitted against a no sanctions alternative.

The Fehr-Schmidt model can also rationalize a "dividend of democracy", i.e. that a scheme chosen by vote may perform better than when imposed exogenously. In particular, voting may serve as an equilibrium selection device. Assume, for example, that a group chooses between no sanctions and non-deterrent formal sanctions, and that multiple, symmetric equilibria exist under the latter scheme. Voting for non-deterrent formal sanctions is only rational if a voter believes her group can coordinate on each member contributing $C'' \geq (sE + c) / (nm + s - 1) = \underline{C}$ to the group account (see Appendix B). Such a belief induces an inequality-averse subject *i* to contribute *C''.* Voting for non-deterrent formal sanctions therefore credibly signals an intention to contribute at least $\underline{C}$ and thereby induces other, inequality-averse group members to do the same. In this way, voting for non-deterrent sanctions induces selection of equilibria with high contributions. The logic is analogous for voting on informal sanctions.[8]

[8] Results are formally derived in Appendix B under the assumption that preferences are common knowledge. Relaxing this assumption opens the avenue for a signaling explanation of a "dividend of democracy". In particular, observing the outcome of the vote may lead subjects to update their beliefs about the prevalence of social preferences in their group which in turn may affect contributions.

In summary, the Fehr-Schmidt model predicts that fixed costs reduce the popularity of formal sanctions since the gain from having them in operation is smaller than predicted by standard theory. We also expect strong support for informal vs. no sanctions, relatively strong support for IS vs. costly and deterrent sanctions, and weaker support for IS vs. cheap and deterrent sanctions. The model also predicts some support for non-deterrent sanctions (when cheap) vs. no sanctions, and a "dividend of democracy" due to signaling which promotes coordination on high-contribution equilibria. However, these predictions hinge on the distribution of social preference parameters and often involve multiple equilibria.

While we find that the Fehr-Schmidt model provides more accurate predictions than standard theory in many dimensions, it is worthwhile to point out that it is an equilibrium model assuming perfect rationality and complete information. However, the empirical facts (punishment and learning occurs) suggest that subjects' ability to coordinate (or knowledge of one-anothers' preferences) is imperfect.[9]

3. EXPERIMENTAL DESIGN

Our core 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. Two additional treatments with exogenous sanctions test for endogeneity effects. (Treatments to test the robustness of our results are discussed in Section 5.)

In all treatments, participants are divided into groups of n = 5 members that remain fixed ("partner matching"). The main treatments have 7 phases, consisting of 4 periods each. Every period, each participant receives an endowment of E = 20 points of experimental currency. He or she decides how to allocate this endowment to a "group account" or a "private account". The total amount in the group account is doubled 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.

[9] Incomplete knowledge of others' preferences can also explain why contributions under no sanctions conditions tend to decline over time (e.g. Fischbacher and Gächter, 2010).

Under *Informal Sanctions* (IS), participants observe what fellow group members have contributed to the group account.[10] They then have the opportunity to reduce the earnings of other group members. 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, in order to make the scheme comparable to IS. 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 by majority vote. In each vote, only two institutions are available for choice, to rule out strategic voting. Voting is simultaneous, and free, and each subject must vote for one of the institutions available (i.e. no abstentions). Subjects learn what scheme was chosen but not the specific number of votes for it.

Figure 1 shows the time line. 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 rule, and the fact that there will be six votes, each governing four periods of play.[11] These instructions also are accompanied by brief oral instructions, control questions, and answering of any questions raised by the subjects. Each of the following phases starts with a vote. In Phase 2, voting is on NS vs. IS, in Phase 3 on NS vs. FS, and in Phase 4 on FS vs. IS. This cycle is repeated in phases 5 to 7.[12] The instructions, included in Appendix A, use neutral language, avoiding terms such as “public good”, “contribute”, or “punishment.”

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

[11] 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 game.

[12] Our robustness treatments avoid order effects by having subjects vote between IS and FS only.

In FS, the four treatments differ by the sanction rate *s* and the scheme's cost *c* (see Endogenous Treatments in Table 1). With *s* = 0.8, formal sanctions are deterrent, while with *s* = 0.4 they are non-deterrent (*s* = 0.6 is the threshold value for zero vs. full contribution, see eq. 2). The fixed cost are *c* = 2 or c = 8. These values correspond to 10% and 40%, respectively, of the hypothetical gains from full cooperation ($\mathscr{P}$), and are referred to as "cheap" versus "expensive" below. 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 learning those of their own session only.

In IS, it costs a sender 1 point to reduce the earnings of the receiver by 4 points; hence σ = 4 (see eq. 3).[13] The following restrictions apply to sanctioning. Each subject is allowed to allocate at most 10 reduction points to each other group member per 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.[14] With these rules, earnings under IS are given by (3'), which modifies (3) using the values of *E* = 20 and σ = 4.

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

Exogenous treatments. To test for a "dividend of democracy", we conduct control treatments in which subjects experience the same sequences of conditions of play but without ever voting on sanction schemes (see the right column of Table 1). Details of the treatments are given in section 4.3 where we report the corresponding tests for effects of endogeneity.

Predictions. Voting predictions according to standard theory are as follows (see section 2 for details): groups are indifferent between NS and IS, always select NS or IS over non-

[13] The 1:4 cost ratio is used elsewhere (e.g. Page *et al*. 2005, Bochet *et al*. 2006, and Nikiforakis and Normann 2008), and the first punishment point purchased in Fehr and Gächter (2000) can cost the recipient more than four times what it costs the sender. Related experiments (e.g. Fehr and Gächter 2002, Egas and Riedl 2008) have used punishment technologies with less power. We check the robustness of our findings to a less effective punishment technology in Section 5.

[14] Both restrictions are common in the literature and are rarely binding in our experiment (about 1% of the cases).

deterrent FS, and always select deterrent FS over NS or IS. Parameter *s* is thus decisive and parameter *c* irrelevant for collective action in all votes involving an FS option. According to the Fehr-Schmidt model of inequality aversion, on the other hand, groups weakly prefer IS over NS. Depending on the distribution of inequality-aversion parameters, some groups may vote for non-deterrent FS over NS and IS. The popularity of FS depends both on deterrence level (*s*) and on fixed cost (*c*). The model predicts FS to be more popular when pitted against NS than when pitted against IS.

Implementation. The experiment was conducted at the Centre for Experimental Economics, University of Copenhagen using the software Z-tree (Fischbacher 2007). We conducted three experimental sessions per treatment (see Table 1). In total, 260 subjects participated in the endogenous treatments, with a further 75 in the exogenous ones (and 255 participants in the robustness treatments to be discussed in section 5). Slightly over half of the 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. At the end of the experiment, each subject's earnings 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.

4. RESULTS

Since our ultimate interest is in the endogenous emergence of institutions, we start with discussing voting outcomes in Section 4.1. Section 4.2 discusses contribution (and in IS also punishment) behaviors and their earnings consequences. Section 4.3 discusses the "dividend of democracy" by comparing play under chosen vs. imposed IS and non-deterrent FS conditions.

4.1. *Voting*

Figure 2 shows the voting outcomes in our four main treatments over time. Results for voting on **NS vs. IS** show that informal sanctions are initially unpopular (about 20 percent of groups accept IS, see Vote 1). But popularity of IS significantly increases with experience (p = .000, Wilcoxon signed-rank test) and IS are chosen by at least 50 percent of groups in the second cycle (see Vote 4). The strong increase in popularity with experience is in line with Ertan *et al*. (2009) and Gürerk *et al*. (2006).

Results for voting on **NS vs. FS** (see Vote 2 and 5 in Figure 2) show that fixed costs of FS matter much for popularity. For example, over both cycles, about three quarters of groups choose deterrent FS when they are cheap, but only about one quarter do when they are expensive. While deterrent FS tend to be more popular than non-deterrent ones (holding costs constant), the fixed-cost effect seems to trump the deterrence effect on popularity. For example, non-deterrent but cheap sanctions are more popular than deterrent but expensive sanctions (when compared to NS), even in the second cycle of voting. The finding that expense matters more to adoption than deterrence is in strong contrast to standard theory but is consistent with the Fehr-Schmidt theory if high β's make for substantial cooperation without sanctions. Increasing popularity of FS in the **DC**, **DE** and **NE** treatments over the cycles suggests that experience of free-riding in earlier phases may have convinced additional groups of the virtue of a sanctions scheme; we consider the effects of such experience in Table 2 below.

Results for voting on **FS vs. IS** (see Vote 3 and 6 in Figure 2) show that except for **DC**, informal sanctions are preferred to formal ones by about 3 out of 4 groups. The ranking of popularity for the various types of FS is similar when FS is pitted against NS or against IS (compare e.g. ranking in Vote 2 and 3), but support for FS is lower when IS is the alternative, as predicted by the Fehr-Schmidt theory. Only in the **DC** treatment are FS more popular than IS, and even there a substantial minority (43%) of groups chooses IS. Only in the **DE** treatment does the share of groups choosing FS rise from Vote 3 to Vote 6, but despite this, support for FS remains strikingly low (about a third of groups).

In summary, the popularity of sanction schemes can be ranked as follows (second cycle, by number of groups): IS $\succ$ NS $\succ$ FS when sanctions are non-deterrent (**NC** and **NE**), IS $\sim$ NS $\succ$ FS in the **DE** treatment, and FS $\succ$ IS $\succ$ NS in the **DC** treatment. With experience, then, informal sanctions become at least as popular as formal ones except when the latter are both deterrent and cheap.[15]

[15] Appendix table C.1 shows the exact vote shares at the group and at the individual level, respectively. Differences between these shares are on the whole relatively small.

We use regression analysis to investigate in more detail how treatment parameters *s* and *c* affect voting. We estimate the following group-level probit model of voting for FS:

$$v_{gT} = 0.5 + \gamma_1 s_g + \gamma_2 c_g + \gamma_3 IS_T + \kappa' X_{gT} + \theta_g + \varepsilon_{gT} \tag{5}$$

$$prob\left(\text{accept FS}\right)_{gT} = prob\left(v_{gT} > 0.5\right) = \Phi\left(\gamma_1 s_g + \gamma_2 c_g + \gamma_3 IS_T + \kappa' X_{gT} + \theta_g\right) \tag{5'}$$

Equation (5) regresses the share of members in group *g* voting for FS in phase *T* on a dummy for deterrence of FS (s_g), a dummy for cheap FS (c_g), and a dummy IS_T which is set to 1 when the choice is FS vs. IS, and to zero when it is FS vs. NS. Depending on the specification, we add a vector of controls for experience in previous phases (X_{gT}). Since s_g and c_g are exogenously and randomly assigned to groups, they are by construction uncorrelated with ε_{gT} and θ_g , which captures unobserved group effects (Φ is the cumulative density function for the standard normal distribution). We assume that θ_g is independently and normally distributed because unobserved heterogeneity in preferences for FS (θ_g)is not correlated with explanatory variables by virtue of random allocation of subjects to groups, and estimate random effects models.

Standard theory predicts (see section 2) that γ_1 is positive while γ_2, γ_3 are zero. The Fehr-Schmidt model predicts that γ_1 and γ_2 are positive and γ_3 is negative.[16] Being simple equilibrium theories, both accounts predict $\kappa = 0$, i.e. that experience is irrelevant. However, experience may matter because it induces individuals who are uncertain about the distribution of preferences to update beliefs.

Table 2 shows that groups are significantly more likely to vote for FS when they are deterrent (by about 30 percentage points) and cheap (by about 45 percentage points). Groups are significantly less likely to vote for FS when the alternative is IS ($\gamma_3 < 0$). These results hold both without (Column 1) or with (Column 2) controls for experience. The observed effects of all three variables are consistent with the results derived from the Fehr-Schmidt model. With controls added, the coefficients on Cheap FS (γ_2) are larger than those on Deterrent FS (γ_1), supporting our result that cost matters more than deterrence.[17]

[16] Appendix B.III shows that the simple version of the Charness-Rabin model makes the same qualitative predictions, except predicting that $\gamma_3 = 0$.

[17] In Model 2, the difference between γ_1 and γ_2 is statistically significant ($p = .082$, Wald test). In Model 1, the same test yields a *p*-value of .107. We separately estimated a regression for the 52

The experience controls show that FS is more likely to be chosen if a previous adoption of FS was successful (i.e. yielded average contributions of more than about 10 points: 10*0.06 – 0.59 > 0), and vice versa. Experience with IS reduces the popularity of FS, but not if informal sanctions were used heavily (i.e. were costly).

High cost-effectiveness explains popularity of informal sanctions (IS)

One of the key results from the discussion above is the striking popularity of IS. Below, we show that the high cost-effectiveness of IS explains this finding. We derive a measure of cost-effectiveness of sanctions (*CE*) to compare the efficiency of alternative sanction schemes. Intuitively, the *CE* measure compares the total costs of sanctions with the social benefits of the higher contributions they help to elicit. Thus, *CE* provides direct information about how profitable a particular sanction scheme is compared to a situation without sanctions. We calculate cost-effectiveness as

$$CE_{ST} = \left(GrossGain_{ST} / Cost_{ST} \right) , \qquad (6)$$

where $GrossGain_{ST} = \Pi_{ST} - \Pi_{NS1}$, i.e. the difference between gross earnings before deduction of sanction costs under regime *S* in phase *T* and earnings in Phase 1 (where a No Sanctions regime was imposed).[18] Gross earnings are $\Pi_{ST} = \sum_{g(S)} \sum_{t=1}^{t=4} \sum_{i=1}^{n} \pi_{igt,ST}$, where *g*(*S*) denotes groups using regime *S* and π_{it} is calculated according to eq. (1) for all treatments. For IS, $Cost_{IST} = \left(1+\sigma\right) \sum_{g(IS)} \sum_{t=1}^{4} \sum_{i=1}^{n} p_{ijgt,IST}$ (with σ = 4 points), and for FS $Cost_{FST} = \sum_{g(FS)} \sum_{t=1}^{4} \sum_{i=1}^{n} \left(c + s\left(E - C_{igt,FST} \right) \right)$, i.e. includes both the fixed and variable cost of sanctions.[19]

observations of the first vote only, between FS and NS, and found that only Cheap FS, not Deterrent FS, obtains a significant coefficient.

[18] Using Phase 1 NS contributions provides a group-specific standard for both the IS and the FS comparisons that is available in all phases, including those when no NS observations are available.

[19] 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 C.1.

Figure 3 shows this cost-effectiveness measure in the first and second voting cycle, by treatment. The upper left panel shows that deterrent and cheap formal sanctions (**DC**) are highly cost-effective, yielding an average return of 2.7 points from increased cooperation per point spent on sanctions in the second cycle. All other types of formal sanctions are not cost-effective on average, i.e. have $CE < 1$. For example, a point spent on non-deterrent and expensive FS (NE) reduces returns from cooperation (by 0.1 point). In contrast, informal sanctions are highly cost effective. The returns from IS increase with experience, and yield clearly higher returns than any of the FS on average (by about a factor of 4 or 5 where this can be calculated).

The results shown in Figure 3 help to explain why IS were so popular with voters but CE is a relatively rough measure in that it reflects the average profitability of a scheme over all groups that used it. As discussed next, a more detailed analysis of how individual earnings shape voting supports the results reported above. We find that an individual is significantly more likely to vote for a sanction regime that was profitable, i.e. had yielded him or her the highest earnings in the past.

We estimate a series of probit regressions (reported in Table C.3 in the Appendix) including only subjects who had previously experienced both schemes under consideration. All of the relative earnings variables obtain coefficients that are highly significant and of the sign consistent with voting for the scheme that gave higher earnings. We obtain these results while controlling for earnings variability and treatments conditions (none of which are significant). While highly suggestive, these regression results need to be taken with a grain of salt because some of the samples are small and the requirement of past experience of both schemes causes unavoidable endogeneity.

4.2. *Contributions and punishment*

This section discusses how the various sanctions schemes affected contributions. In particular, we show that if IS was chosen, informal sanctions were behaviorally deterrent in that IS were so well-targeted that contributing was profitable for a rational subject.

Figure 4 shows average contributions over time by treatment. In each treatment, groups operating under a no sanctions regime—the standard VCM—display patterns familiar from other experiments: the average contribution begins at 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. Such contributions, while inconsistent with standard theory predictions, are not inconsistent with the Fehr-Schmidt model in general but *are* higher than predicted by Fehr and Schmidt's estimates of the empirical distribution of inequality aversion parameters.[20]

The upper two panels of Figure 4 show that deterrent formal sanctions lead to contributions between 80 and 100% of endowment (see lines marked with triangles). With full efficiency never quite achieved by deterrent FS, and with NS play showing average contribution levels considerably above zero, the "cooperation premium" $\mathscr{P}$ is clearly smaller than standard theory indicates.[21] The lower two panels (lines with triangles) show that non-deterrent sanctions induce higher contributions than NS, in contrast to the predictions of standard theory.[22]

Informal sanctions induce contribution levels similar to deterrent sanctions (compare lines with squares and triangles in the upper panels). This is inconsistent with standard theory which predicts that the opportunity to impose IS has no effect, but is consistent with the Fehr-Schmidt model assuming enough agents with high α**.** Contributions under IS are similarly high in **NC**, and therefore exceed those under (non-

[20] See Appendix B.I for computations of equilibria with empirical parameters presented in Fehr and Schmidt (1999). Contributions may be higher than predicted by both standard and Fehr-Schmidt theory because agents are acting in a non-equilibrium environment, lacking knowledge of one another's preferences. Initially high contributions, their gradual decay, and their rise again in later periods, are also consistent with the idea of conditionally cooperative preferences (Fischbacher and Gächter, 2010) and with that of signaling effects from voting.

[21] Standard theory predicts $\mathscr{P} = E(mn\text{-}1) = 20$ and a net gain of 12 (= $\mathscr{P}$ - c) in DE. When DE is available but not chosen, observed average contributions are about 12, and earnings are 32 with all contributing 12 in NS. When DE is chosen, the average contribution is about 19 (95% of the endowment), and subjects earn 39 with all contributing 19 *before* subtracting the cost *c*. Thus, the fixed cost in DE of $c = 8$ is in fact very close to the cut-off point for DE to be profitable under actual behaviors. Less than full contribution with deterrent sanctions, perhaps due to error or a dislike of being coerced, also reduces $\mathscr{P}$ but is not predicted by any of the social preference theories considered here.

[22] Contributions are higher with non-deterrent sanctions in **NC** and **NE** than with NS in the second cycle ($p < 0.05$, Mann-Whitney tests for phase 6). See Appendix Table C.4 for details.

deterrent) FS there. In **NE**, contributions under IS average roughly 70% of the endowment, resembling average contributions under FS but clearly exceeding those in NS condition.[23]

Informal sanctions were meted out moderately and mostly with circumspection. For example, subjects purchased an average of 1 punishment point per period under IS, and the average recipient of punishment was targeted with 3 points of punishment and thus lost 12 points. As shown below (and in Appendix Table C.2) in more detail, sanctions were mostly targeted at low contributors. A free-rider who is targeted with at least 12 punishment points prefers fully contributing to free riding. Full free ridings yields an income of 20 - 12 + 0.4 ΣC_{-i}, full cooperation 0 + 0.4 $\cdot$ 20 + 0.4 ΣC_{-i} . Thus, informal sanctions were "behaviorally deterrent" because they were well-targeted and sufficiently abundant. As a result, we find that subjects tended to earn more by contributing more (result from GLS regressions of earnings on contributions in the IS condition are reported below Table C.2 in the Appendix).

Table 3 shows that informal punishment was mostly well-targeted at free riders, and that it had a disciplining effect, i.e. induced higher contributions. Regressions (1) and (2) serve to explain the determinants of punishment at the individual level, i.e. of subject *j* by subject *i* (random effects GLS regressions, Tobit regressions yield the same qualitative results). First, subjects who contribute more get less punishment (see negative coefficients on $C_{j,t}$). In addition, relative contributions by *i* and *j* matter. Subject *j* gets more punishment by *i* if *j* contributed less than *i* (see coefficients on the variable with the min operator), as predicted by the Fehr-Schmidt model. To be sure, we also find evidence of perverse punishment in which *i* punishes a subject that contributed more than *i* did. But the latter coefficients tend to be smaller than the former, indicating more pro-social than anti-social punishment. Finally, the insignificant coefficients on Voted for IS_j indicate that those who voted for IS do not punish more. These results hold whether or not we control for group-specific effects by adding group dummies.

Regressions (3) and (4) serve to explain the effects of informal sanctions on subsequent cooperation behavior. We find, first, that low contributors are disciplined by punishment. The further they were below the group median, and the more punishment they

[23] Contributions are higher with IS than with NS or with non-deterrent FS ($p < 0.05$, MW tests). We find no difference when deterrent sanctions are used in **DC** and **DE**. See Appendix Table C.4 for details.

got, the stronger the subsequent increase in contributions (see line 6 in Table 3). Note that the coefficients are highly significant despite controlling for own contributions in the previous period (see first line). High contributors, if anything, tended to be demotivated by punishment. The controls for own and group average contribution also obtain positive coefficients (significant in all cases except group average contributions in model 4), indicating persistence of own tendency to contribute and reinforcement by other group members' contributions. Subjects who voted in favor of IS do not contribute significantly more than those who voted against.

4.3. *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 in NS for experienced subjects (i.e. in Phase 5).[24]

Based on previous research (Tyran and Feld 2006, Dal Bo *et al*. 2010, Sutter *et al*. 2010) we speculate that voting on sanctions creates a "dividend of democracy," i.e. that the high efficiency of IS (and to a lesser extent non-deterrent sanctions) in our experiment was partly due to the fact that it was chosen by the subjects over some alternative. For example, subjects in the **DC** treatment who learn that the group voted for IS 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. More generally, votes for IS might coordinate beliefs to select equilibria with high contributions and thereby also reduce punishment stemming from coordination failure. So an IS vote outcome might lead to higher contributions through presumed signals of intent to cooperate and punish (cf. the discussion in Section 2).

[24] Average earnings under NS and IS are not significantly different in phase 2, with the exception of the **NC** treatment (p = .07). 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). All tests in this note are two-tailed Mann-Whitney tests at group level. Within-group tests for cases in which a given group can be observed under both IS and NS in different phases give similar results.

We test this conjecture by exogenously imposing the same sequence of rules that was endogenously chosen in treatment **DC**. In the **Exogenous IS** treatment, subjects face the rules experienced by counterparts in the **DC** treatment in the order that was the most common path leading to a trial of IS in that treatment (NS in phases 1 and 2, FS in Phase 3, IS in Phase 4). This sequence occurred endogenously for 4 groups, and we have 6 groups experiencing the same rules in **Exogenous IS** but with no mention of voting.[25] Notice that in both **DC** and **Exogenous IS** subjects face 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 strong evidence of a dividend of democracy for the informal sanctions regime. Average contributions and earnings are about 30 percent higher when IS is chosen than when it is exogenously imposed, holding experience constant. Both differences are significant in two-tailed Mann-Whitney tests at the group level (average contributions in Phase 4: 14.8 vs. 18.8, $p < 0.05$, earnings: 27.7 vs. 35.6, $p < 0.05$). These results are fully consistent with our conjecture that IS performs better when chosen by vote than when assigned exogenously (we discuss selection issues in the footnote).[26]

We proceed analogously to test for a dividend of democracy with non-deterrent formal sanctions. In the **Exogenous Non-deterrent FS treatment**, we exogenously impose the most common order seen in the **NC** treatment (NS in Phases 1 and 2, non-deterrent and cheap formal sanctions in Phase 3). We again find supportive evidence for our conjecture,

[25] Instructions and procedures were identical to **DC**, except that subjects were told that the computer would decide which rule they would be assigned, and were not told what that decision would be based on. We chose to implement the most commonly observed path leading to an IS condition in **DC** to maximize the 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. Note that we only compare groups following the exact same institutional path up to Phase 4.

[26] A possible concern is that pro-IS voters are intrinsically more cooperative than pro-FS voters and that IS thus performs better because cooperative subjects tend to self-select into that condition. We find no significant support for this concern. In many cases when IS is accepted, some have voted for IS and some against IS (i.e. are pro-FS voters). But pro-IS voters do not contribute significantly more than pro-FS voters (19.2 vs. 18.6, $p = .12$, MW test). Neither are pro-IS voters more cooperative in phase 1 (when NS is imposed) than pro-FS voters (12.5 vs. 10.8, $p = .13$, MW test).

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 groups using those sanctions endogenously following the same 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 (on selection issues, see the footnote).[27] Since earnings are perfectly correlated with contributions in FS, the difference in average earnings is significant at the same level.

5. ROBUSTNESS CHECKS

One of our most surprising results is the relatively high popularity of IS, even when pitted against deterrent FS. We test for robustness along two dimensions. First, we investigate whether increasing the cost of punishment (i.e. the fee-to-fine ratio 1:σ) reduces the effectiveness and, hence, the popularity of informal sanctions. Second, we check whether our design choice to implement an "either-or" choice between IS and FS biased results by failing to allow the two types of sanctions to coexist. The main treatments may have given deterrent FS a best shot since adopting FS meant suppressing IS, but one might argue that this design choice does not parallel all natural settings where IS may continue to operate when FS are introduced.[28] Both of these checks are performed under essentially identical conditions as in the main treatments, except that we allow for more learning by implementing six rather than just two votes between FS and IS.[29] As a control, we also conduct two treatments with the punishment effectiveness rate (σ) used in the main treatment, without coexistence and with 6 votes.

[27] We check for self-selection following the same logic as with IS (see previous footnote) and, again, find no support for selection. If non-deterrent sanctions are in place, pro-FS voters do not contribute significantly more than anti-FS voters in Phase 3 (14.0 vs. 14.5, *p* = 0.73, MW test). Members of groups choosing FS endogenously also did not contribute more in phase 1 than those in the exogenous FS treatment (10.1 vs. 9.2, *p* = .48, MW test).

[28] Suppression of severe forms of decentralized punishment (vendettas, vigilante action etc.) is a fundamental task for centralized authorities, such as the state. In this sense, an authority with the ability to mete out formal sanctions tends to come with a reduction in decentralized punishment. On the other hand, milder forms of informal sanctions (shunning, bad-mouthing etc.) are never fully suppressed by a centralized authority.

[29] We also drop the initial phase 1 in which the condition without sanctions (NS) is imposed.

Table 4 provides an overview of the treatments checking for robustness. In all treatments, IS and FS are directly pitted against each other six times (rather than just twice as in the main experiments), and subjects never experience a condition without sanctions. We implement deterrent FS only, but for both levels of fixed cost. For example, **DC6** indicates that deterrent formal sanctions are cheap (c = 2) and that there are 6 votes between FS and IS. The main treatments had a fee-to-fine ratio 1:σ of 1:4, and the controls testing for the effects of reduced punishment effectiveness of 1:2 are labeled **DC6_1:2** and **DE6_1:2**. In the treatments with coexistence (**DC6_coex** and **DE6_coex**), the vote is on whether to add FS to IS. If so, subjects are given the opportunity to reduce the earnings of other group members after they have paid formal fees and fines.[30] The numbers in parentheses indicate the number of groups per treatment (n = 255 in total).[31]

Table 5 shows the share of groups voting for FS, by treatment and phase. Voting outcomes of the main experiment for IS vs. FS in the **DC** and **DE** treatments are also included for ease of comparison (see first two columns). The last four columns show that results in treatments with coexistence and with reduced punishment effectiveness in IS are similar to those in the main experiment: FS is much more popular when cheap than when expensive (by about a factor of 4), but IS are remarkably popular overall. If the alternative is deterrent and expensive FS, a vast majority prefers IS (83.3 and 87.5 percent respectively). But support for IS is substantial even when deterrent sanctions are cheap and voters have much opportunity to learn. For example, in the 6th vote, between one third (in **DC6_1:2**) and one half (in **DC6_coex**) vote for informal sanctions when they are predicted by standard theory to obtain zero support.

The control treatments for repetition (**DC6** and **DE6**) show that deterrent and expensive FS do not get more popular with more experience. In fact, in the 6th vote of **DE6**, only 25 percent of groups support FS (we found the same support in **DE** of our main treatments, see line Total). However, support for deterrent and cheap FS is surprisingly low

[30] Kube and Traxler (2009) study combined FS+IS in a setting in which FS is non-deterrent and the scheme is imposed exogenously, while subjects in one treatment in Andreoni and Gee (2012) can add a deterrent formal sanction on the lowest contributor, with IS remaining available.

[31] The sessions were conducted at the same site with students from the same university who were all inexperienced in public goods experiments. Subject characteristics were similar except that the percentage of freshman economics students was smaller, at 16% versus 51%.

in the **DC6** treatment. It is only about half of the support we found for such sanctions in the corresponding main treatment (27 vs. 57 percent). This result strengthens the view that IS is surprisingly popular, but somewhat weakens the view that popularity of FS is strongly influenced by its cost (in **DE6**, 31 percent of groups choose FS).[32]

We do not observe strong learning effects. In particular, there is no clear tendency for FS to become more popular over time in the robustness treatments. This finding mirrors the pattern in the main experiment where popularity of FS vs. IS did not change from first to second voting cycle. It is noteworthy, however, that IS is quite popular in all robustness treatments even in Phase 1. This finding contrasts with earlier studies of voting between IS and NS, which find that IS initially unpopular (e.g. Gürerk *et al*., Ertan *et al*.). But the finding strengthens our overall result that IS is surprisingly popular when competing with FS.

Table 6 presents results from regression analyses supporting our earlier interpretations. They show, in particular, that formal sanctions are less popular when (moderately) expensive, that doubling the cost (equivalently, halving the effectiveness) of informal sanctions does not make formal sanctions more popular, and that these results are not attributable to lack of coexistence of IS with FS or to insufficient opportunities for learning. Model (1) uses data from the robustness treatments only, model (2) also uses the relevant data from the main experiment (voting on IS vs. FS in **DC** and **DE**) and includes a dummy for having six votes between IS and FS. Addition of controls for group experience in specification (3) leaves qualitative conclusions unchanged.[33]

[32] We think the small number of groups voting for FS in **DC6** is best treated as an outlier. Among other things, the difference between the 57% share of groups choosing FS in **DC** and the 27% share in **DC6** could have been undone by a few voters, since four **DC6** groups chose IS in their first vote by a narrow 3-2 majority. Once IS was chosen by a group, its high effectiveness could easily have discouraged voters from experimenting with the alternative.

[33] Figure C.2 in the online appendix shows contributions- and punishment behavior over time. In line with other studies (e.g. Nikiforakis and Normann 2008), we find that lower punishment effectiveness reduces contributions under IS somewhat, although the difference is not statistically significant. We find no evidence that coexistence significantly affects contributions compared to the situation where either institution exists in isolation.

6. CONCLUDING REMARKS

The voluntary provision of public goods is beset by the notorious free rider problem. An obvious remedy for the resulting inefficiency is to sanction free riders. While informal sanctions have been extensively studied in the literature, economists have not paid much attention to the behavioral impact of formal sanctions. The comparative performance of formal and informal sanctions, and their respective popular support, have not been studied previously at all. A plausible reason for this neglect is that standard economic reasoning suggests that well-targeted deterrent formal sanctions are dominant, and will therefore be the voters' preferred institutional choice. This paper has shown that this presumption is overly simple, if not wrong.

Our study confirms that deterrent formal sanctions increase cooperation, in line with standard theory. In contrast to standard theory, but in line with Fehr and Schmidt (1999) and recent experimental evidence, we find that informal sanctions also induce high levels of cooperation. Because our subjects use informal sanctions diligently and in a well-targeted manner, they are "behaviorally deterrent" and therefore increase efficiency. 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 that do not require a costly sanctioning infrastructure, and increasingly vote for them. As a result, informal sanctions are relatively popular, and self-organization for collective action is remarkably successful. We show that these findings are robust to providing more learning opportunities, to reducing the effectiveness of informal sanctions, or to allowing informal and formal sanctions to co-exist.

Our second main finding is that endogenous institutional choice carries a "dividend of democracy". We find significantly greater effectiveness of informal sanctions when selected by voting than in our exogenous comparison treatment. This result appears not to be an artifact of selection effects, and it is in line with others regarding the role of coordination in using punishment effectively, as well as with Dal Bó *et al.*'s (2010) and Sutter *et al.*'s (2010) findings about the benefits of democratic choice of institutions.

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, enhance cooperation when selected democratically, reinforcing the result in Tyran and Feld

(2006). However, even democratically legitimized non-deterrent formal sanctions show less power to induce cooperation than do informal ones (which, in our data, are "behaviorally deterrent", i.e. make contributing privately optimal).

We think our results provide at least four directions for further research. First, our observation of high cost-effectiveness of deterrent sanctions seems to suggest that they are preferable to non-deterrent formal sanctions. But our experiment remains silent as to whether they are in fact preferred by voters, because it provides them only with a choice between formal vs. informal sanctions but not between deterrent vs. non-deterrent formal sanctions.[34] Voters may in fact be reluctant to vote for fully deterrent sanctions because they may require a severity of punishment that would be judged unacceptable, in part due to the possibility of judicial and enforcement errors, an additional topic for research in its own right.

Second, we identify fixed cost as a fundamental qualitative difference between formal and informal sanction regimes: having access to formal sanctions requires putting in place some costly infrastructure. Our experiment shows that high fixed costs undermine popular support for formal sanctions. This finding suggests that keeping such costs low is important for policy makers who seek popular support for implementing formal sanctions. To what extent this is feasible probably depends on the particular application one has in mind. For example, fixed costs might be low if the overall cost of the state enforcement machinery can be spread over a great many domains. An interesting question for future research is to test to what extent popular support for informal sanctions varies as fixed costs of formal sanctions approach very low levels.

Third, an interesting question to investigate is how our finding that informal sanctions are surprisingly popular even when pitted against deterrent formal sanctions maps into larger groups and electorates. One might expect informal sanctions to be less powerful in large groups because a given individual may only be able to observe and punish a subset of that group (Carpenter 2007 finds that group size is relatively unimportant when comparing groups of 5 and 10 persons, but applicability to much larger social groups remains an open question). However, the effectiveness of formal sanctions may also depend

[34] See Kamei *et al*. (2011), which allows group choice of sanction parameters and obtains results consonant with ours.

on the size of the group. For example, fixed costs for the sanctioning infrastructure may be subject to (dis-)economies of scale. Finally, voting outcomes may be shaped by the size of the electorate because voters may realize that they are unlikely to be pivotal in a large electorate and may therefore vote expressively (e.g. Tyran 2004 or Feddersen *et al*. 2009).

Fourth, in essence, we find that informal sanctions are popular because they are effective, and they are effective because they are used with circumspection. But such circumspection, in particular refraining from harmful perverse punishment, and the trust that other group members exhibit such circumspection might depend on norms of civic cooperation or "culture" more generally (see Herrmann *et al*. 2008). But cultural factors may also affect the effectiveness and popularity of formal sanctions. The "rule of law" and willingness to comply with formal rules (e.g. the tax code) seem to vary systematically across countries. It would thus be interesting to conduct our experiments in countries that are believed to have weaker norms of civic cooperation or rule of law.

At a higher level of abstraction, informal and formal sanctions should probably not be viewed entirely as alternatives. Rather, the centralized organizational structures that make formal sanctions a possibility 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 individuals have to cooperate. And cooperation is required to create and sustain the administrative machineries that make formal sanctions an option.

REFERENCES

ANDREONI, J. and GEE, L.K. (2012), "Gun for Hire: Delegated Enforcement and Peer Punishment in Public Goods Provision," *Journal of Public Economics*, **96**, 1036–1046.

BLANCO, M., ENGELMANN, D., and NORMANN, H.T. (2011), "A Within-Subject Analysis of Other-Regarding Preferences," *Games and Economic Behavior,* **72,** 321-338.

BOCHET, O., PAGE, T. and PUTTERMAN, L. (2006), "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior and Organization,* **60**, 11-26.

BOTELHO, A., HARRISON, G., COSTA PINTO, L.M. and RUTSTRÖM, E.E. (2005), "Social Norms and Social Choice," unpublished paper, Dept. of Economics, University of Central Florida.

BUCHANAN, J. and TULLOCK, G. (1962), *The Calculus of Consent. Logical Foundations of Constitutional Democracy.* (Ann Arbor: University of Michigan Press).

CARPENTER, J. (2007), "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods," *Games and Economic Behavior,* **60**, 31-52.

CHARNESS, G. and RABIN, M. (2002), "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, **117**, 817-869.

CHAUDHURI, A. (2010), "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature," *Experimental Economics,* **14**, 47-83.

CINYABUGAMA, M., PAGE, T. and PUTTERMAN, L., 2006, "Can Second-Order Punishment Deter Perverse Punishment?" *Experimental Economics* 9(3): 265-279.

DAL BÓ, P., FOSTER, A. and PUTTERMAN, L. (2010), "Institutions and Behavior: Experimental Evidence on the Effects of Democracy," *American Economic Review,* **100**, 2205–2229.

DENANT-BOEMONT, L., MASCLET, D. and NOUSSAIR, C. (2007), "Punishment, Counter-punishment and Sanction Enforcement in a Social Dilemma Experiment", *Economic Theory,* **33**, 145-167.

EGAS, M. and RIEDL, A. (2008), "The Economics of Altruistic Punishment and the Maintenance of Cooperation," *Proceedings of the Royal Society* B, **275**, 871-878.

ERTAN, A., PAGE, T. and PUTTERMAN, L. (2009), "Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free-Rider Problem," *European Economic Review,* **53**, 495-511.

FEDDERSEN, T., GAILMARD, S. and SANDRONI, A. (2009), "Moral Bias in Large Elections: Theory and Experimental Evidence," *American Political Science Review,* **103**, 175-192.

FEHR, E. and GÄCHTER, S. (2000), "Cooperation and Punishment in Public Goods Experiments,“ *American Economic Review,* **90**, 980-994.

FEHR, E. and GÄCHTER, S. (2002), "Altruistic Punishment in Humans," *Nature,* **415**, 137-140.

FEHR, E. and SCHMIDT, K. (1999), "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics,* **104**, 817-868.

FISCHBACHER, U. (2007), "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics,* **10**, 171-178.

FISCHBACHER, U. and GÄCHTER, S. (2010), "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments," *American Economic Review,* **100**, 541-556.

GÜRERK, Ö., IRLENBUSCH, B. and ROCKENBACH, B., (2006), "The Competitive Advantage of Sanctioning Institutions," *Science*, **312**, 108-110.

HERRMANN, B., THÖNI, C. and GÄCHTER, S. (2008), "Antisocial Punishment Across Societies," *Science*, **319**, 1362-1367.

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

ISAAC, R. M. and WALKER, J.M. (1988), "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism," *Quarterly Journal of Economics*, **103**, 179-199.

KAMEI, K. (2011), "Democracy and Resilient Pro-Social Behavioral Change: An Experimental Study," unpublished paper, Brown University.

KAMEI, K., PUTTERMAN, L. and TYRAN, J.-R. (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.

KOCHER, M.G., MARTINSSON, P. and VISSER, M. (2008), "Does Stake Size Matter for Cooperation and Punishment?" *Economics Letters*, **99**, 508-511.

KOSFELD, M., OKADA, A. and RIEDL, A. (2009), "Institution Formation in Public Goods Games," *American Economic Review*, **99**, 1335-1355.

KROLL, S., CHERRY, T.L. and SHOGREN, J.F. (2007), "Voting, Punishment and Public Goods," *Economic Inquiry*, **45**, 557-570.

KUBE, S. and TRAXLER, C. (2011), "The Interaction of Legal and Social Norm Enforcement," *Journal of Public Economic Theory,* **13**, 639-660.

LOCKE, J. (2005) [1689], *Two Treatises of Government and a Letter Concerning Toleration.* (Digireads.com Publishing, Stilwell).

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

NIKIFORAKIS, N. and NORMANN, H. (2008), "A Comparative Statics Analysis of Punishment in Public Goods Experiments," *Experimental Economics*, **11**, 358-369.

OSTROM, E. (2010), "Beyond Markets and States: Polycentric Governance of Complex Economic Systems," *American Economic Review*, **100**, 641-672.

PAGE, T., PUTTERMAN, L. and UNEL, B. (2005), "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency," *Economic Journal*, **115**, 1032-1053.

SMITH, A. (2003) [1776], *An Inquiry into the Nature and Causes of the Wealth of Nations*. (New York: Bantam/Dell).

SUTTER, M., HAIGNER, S. and KOCHER, M. (2010), "Choosing the Stick or the Carrot? – Endogenous Institutional Choice in Social Dilemma Situations," *Review of Economic Studies*, **77**, 1540-1566.

TRAULSEN, A., RÖHL, T. and MILINSKI, M. (2012), "An Economic Experiment Reveals that Humans Prefer Pool Punishment to Maintain the Commons", *Proceedings of the Royal Society B*, **279**, 3716-3721.

TYRAN, J.-R. (2004), "Voting when Money and Morals Conflict. An Experimental Test of Expressive Voting," *Journal of Public Economics*, **88**, 1645-1664.

TYRAN, J.-R. and FELD, L.P. (2006), "Achieving Compliance when Legal Sanctions are Non-deterrent," *Scandinavian Journal of Economics*, **108**, 1-22.

TYRAN, J.-R. and SAUSGRUBER, R. (2006), "A Little Fairness may Induce a Lot of Redistribution in Democracy," *European Economic Review,* **50**, 469-485.

ZHANG, B., LI, C., DE SILVA, H., BEDNARIK, P. and SIGMUND, K. (2013), "The Evolution of Sanctioning Institutions: An Experimental Approach to the Social Contract." Working Paper, University of Vienna.

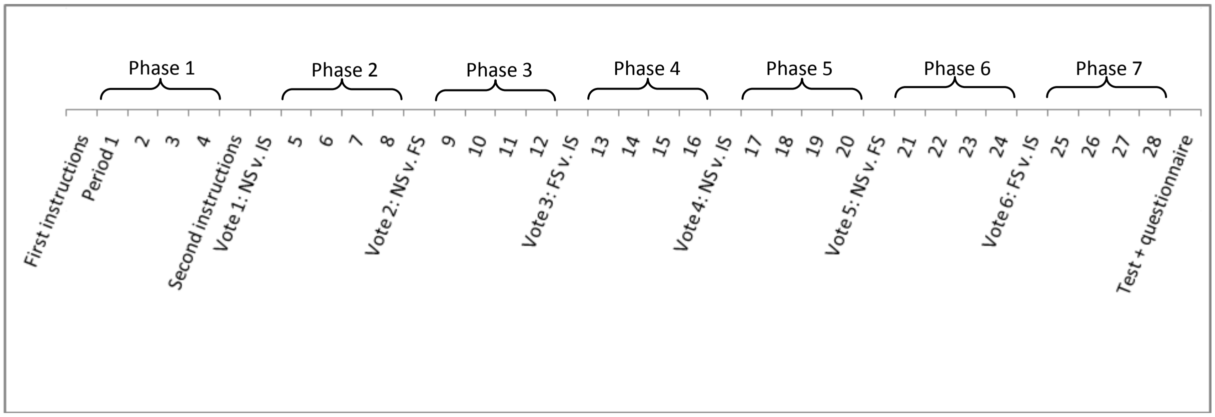


FIGURE 1

Timing in the experiment

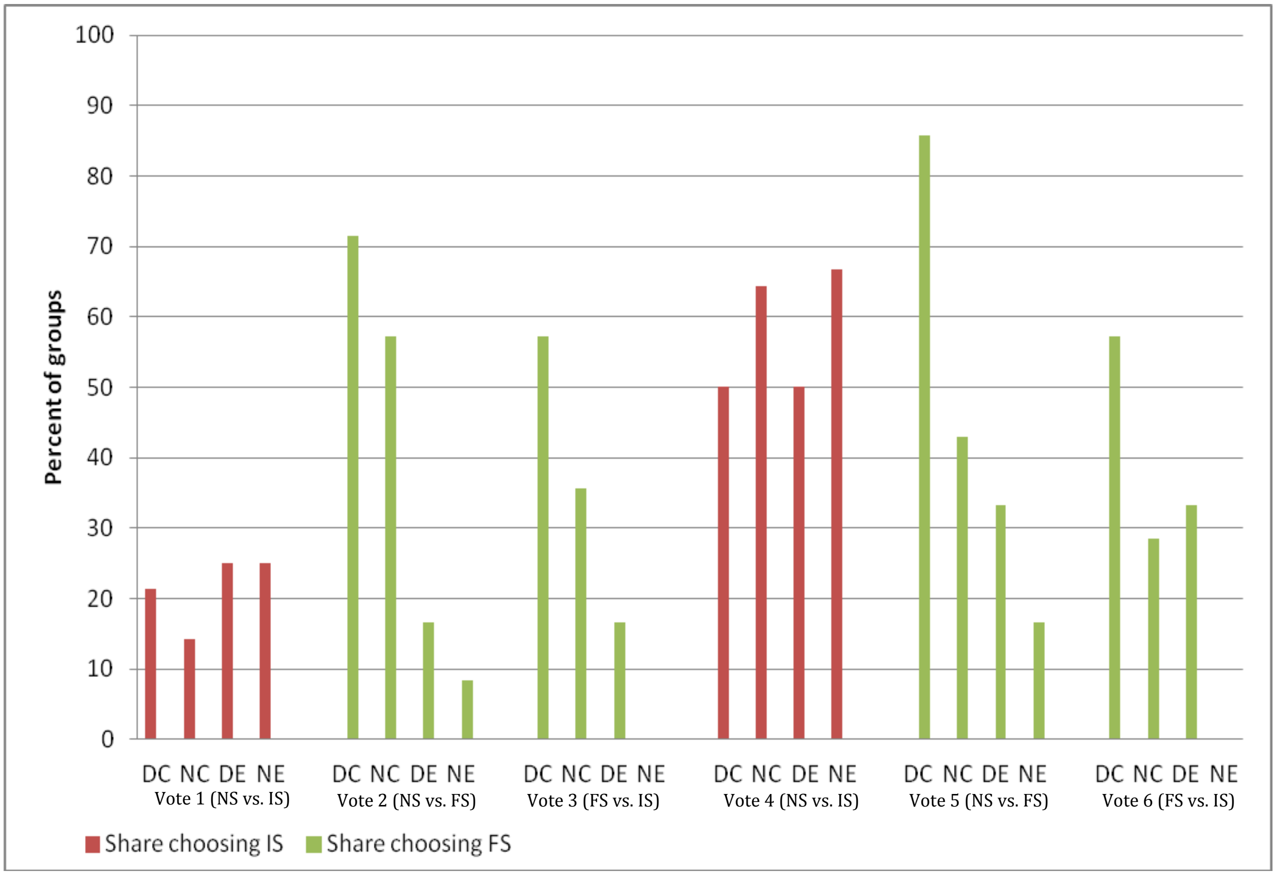
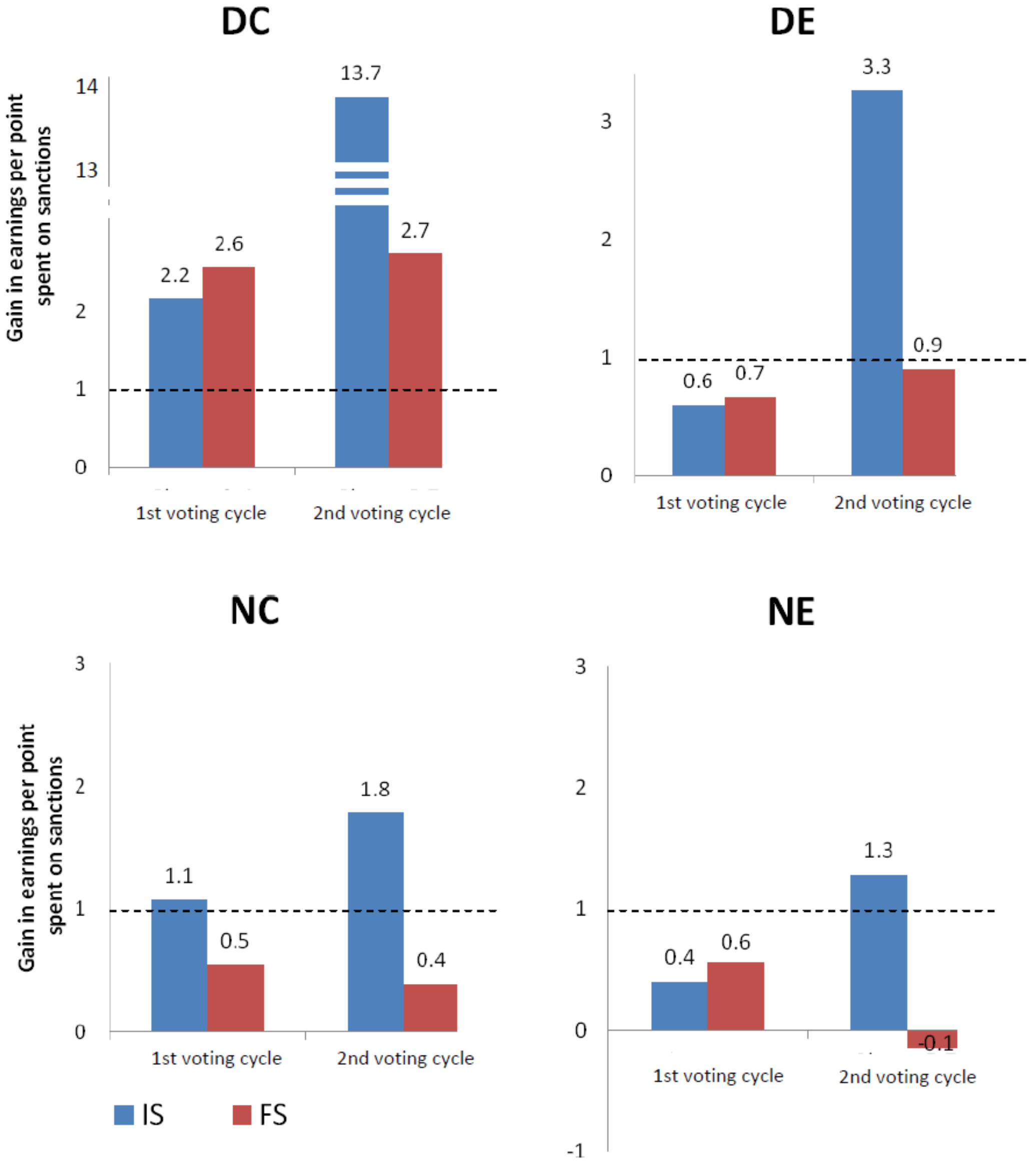


FIGURE 2

Voting outcomes



Note: Bars show *CE*, the average gain in gross earnings (i.e. before sanction costs are deducted), relative to earnings in Phase 1, for each point spent on sanctions, including losses to those punished (see main text for an exact definition). Earnings in Phase 1 are calculated separately for the groups that experienced the relevant institutions in the relevant phases. Dashed line at *CE* =1 indicates the break-even point, at which earnings gain equals sanction cost. Note that the scale of the *y*-axis is different in the upper left (**DC)** quadrant.

FIGURE 3

Cost-effectiveness of sanctions

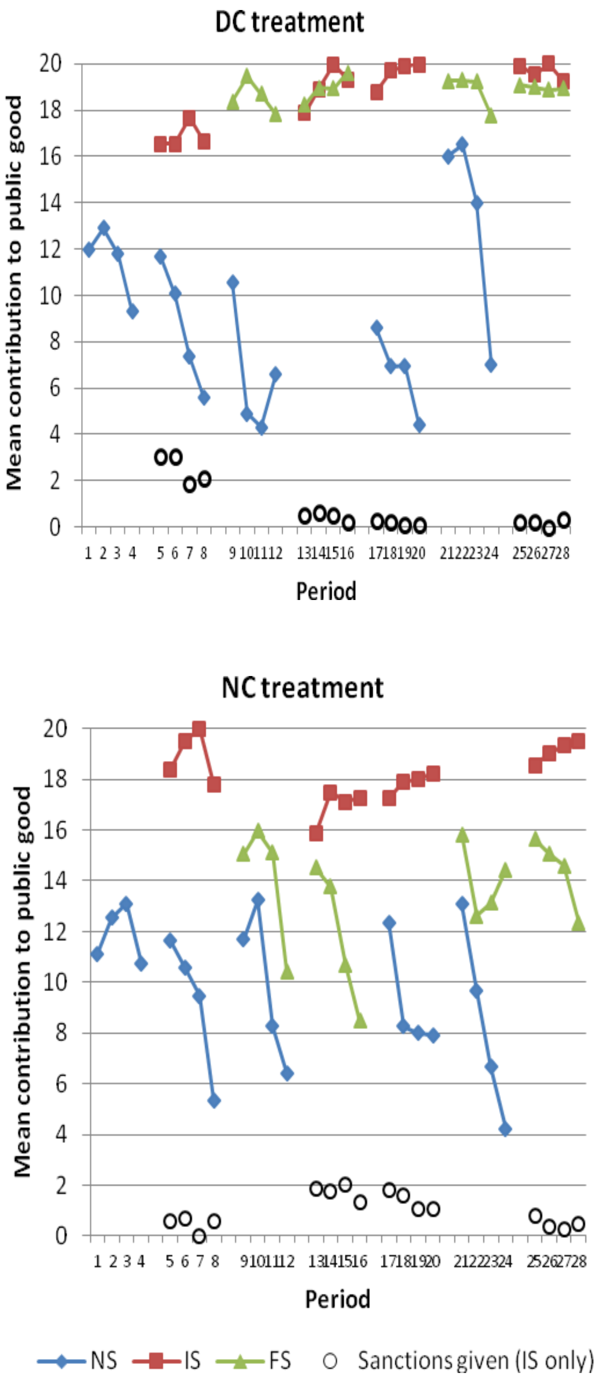


FIGURE 4

Contributions

TABLE 1

Main Treatments (Formal Sanctions Parameters)

	Endogenous Treatments		Exogenous Treatments *c* = 2
	c = 2 "Cheap"	*c* = 8 "Expensive"	
s = 0.8 "Deterrent"	Deterrent and Cheap (**DC**) Groups: 14	Deterrent and Expensive (**DE**) Groups: 12	**Exogenous IS** (Informal Sanctions) Groups: 6
s = 0.4 "Non-deterrent"	Non-deterrent and Cheap (**NC**) Groups: 14	Non-deterrent and Expensive (**NE**) Groups: 12	**Exogenous Non-deterrent FS** (Formal Sanctions) Groups: 9

Note: All groups had 5 members. Main treatments: 52 groups with a total of 260 subjects. Exogenous treatments: 15 groups with 75 subjects.
In both exogenous treatments, *c* = 2 while the value of *s* is as indicated by the relevant row heading. For treatments testing for robustness, see Table 4.

TABLE 2

Voting regressions, group level

	Dep. Var.: group adopted formal sanctions	
	(1)	(2)
Deterrent FS (γ_1)	0.301***	0.238**
	(0.091)	(0.105)
Cheap FS (γ_2)	0.471***	0.446***
	(0.082)	(0.117)
Alternative to FS is IS (γ_3)	-0.205**	-0.407***
	(0.101)	(0.116)
FS used before		-0.590*
		(0.173)
Contributions * FS used before		0.060**
		(0.028)
IS used before		-0.739*
		(0.243)
Contributions * IS used before		0.025
		(0.027)
Mean informal punishment given before		0.125**
		(0.057)
Mean contribution in *T*-1		-0.002
		(0.010)
Mean contribution in *T*-2		-0.016
		(0.012)
Phase dummies	Yes	Yes
Log-likelihood	-105.3	-90.2
Observations	208	208
Number of groups	52	52

Note: Random effects probit regressions. Standard errors in parentheses. Marginal effects reported. Unit of analysis: groups. Data from all four phases where subjects vote on introducing FS. "FS used before" = 1 if a group implemented FS in the most recent opportunity to do so. "Mean informal punishment given before" is the average number of punishment points meted out by all group members in the most recent phase where IS were available. "Contributions" are group averages. *T*-1 indicates the phase immediately preceding the vote in phase *T*. In interactions, both terms always refer to the same phase, i.e. if IS was used in Phase *T*-1, both terms in "Contributions*IS used before" refer to Phase *T*-1. Constant included (not shown).

* significant at 10%; ** significant at 5%; *** significant at 1% (refer to the test of the hypothesis that the coefficient is zero)

TABLE 3

Determinants of punishment and contributions under IS

	Dependent variable:			
	Punishment points given to individual other		Contribution	
	(1)	(2)	(3)	(4)
$C_{j,t-1}$			0.587***	0.511***
			[0.054]	[0.071]
$\bar{C}_{-j,t-1}$			0.312***	0.005
			[0.050]	[0.059]
$C_{j,t}$	-0.008	-0.010**		
	[0.011]	[0.004]		
$\max\left(C_j - C_i, 0\right)$	0.039***	0.039***		
	[0.013]	[0.003]		
$\left\|\min\left(C_i - C_j, 0\right)\right\|$	0.095***	0.092***		
	[0.011]	[0.005]		
$\text{Punishment received}_{t-1} * \left\|\min\left(C_{j,t-1} - C_{m,t-1}, 0\right)\right\|$			0.031***	0.029***
			[0.010]	[0.010]
$\text{Punishment received}_{t-1} * \max\left(C_{j,t-1} - C_{m,t-1}, 0\right)$			-0.047	-0.024
			[0.029]	[0.026]
Voted for IS$_j$	-0.014	-0.014	0.299	0.279
	[0.040]	[0.023]	[0.248]	[0.261]
Group dummies	No	Yes	No	Yes
R-sq (overall)	0.21	0.25	0.62	0.65
Observations	9,120	9,120	1,710	1,710
Number of dyads	940	940		

Note: Random effects GLS regressions. Standard errors in parentheses. Standard errors clustered by group. Units of observation are dyads-by-period in the first two regressions and individuals-by-period in the last two. Only observations in IS condition are included. $C_{j,t}$ is the contribution to group production by individual *j* in period *t*. $C_{m,t}$ is the median contribution in period *t*. A constant, controls for phase, period within phase, and treatment are included (not shown). * significant at 10%; ** significant at 5%; *** significant at 1%

TABLE 4

Robustness treatments

		Coexistence between IS and FS	
Punishment effectiveness in IS	**Cost of FS**	*No*	*Yes*
1:4	*Cheap*	DC6 (8)	DC6_coex (8)
	Expensive	DE6 (8)	DE6_coex (8)
1:2	*Cheap*	DC6_1:2 (9)	
	Expensive	DE6_1:2 (10)	

Note: Number of groups in parentheses. All groups had 5 subjects. Total number of participants: 255

TABLE 5

Voting outcomes at the group level, robustness treatments

	Treatment							
Vote	DC	DE	DC6	DE6	DC6_1:2	DE6_1:2	DC6_coex	DE6_coex
1			12.5	50.0	55.6	20.0	0.0	12.5
2			37.5	50.0	88.9	20.0	62.5	12.5
3	57.1	16.7	25.0	25.0	77.8	0.0	87.5	0.0
4			37.5	12.5	66.7	20.0	75.0	12.5
5			37.5	25.0	55.6	20.0	50.0	25.0
6	57.1	33.3	12.5	25.0	66.7	20.0	50.0	12.5
Total	57.1	25.0	27.1	31.3	68.5	16.7	54.2	12.5

Note: Entries are the share of groups choosing formal sanctions, in percent.

TABLE 6

Voting regressions, group level, robustness treatments

	Dependent variable: Adopted formal sanctions		
	(1)	*(2)*	*(3)*
Cheap FS	0.378***	0.401***	0.353***
	(0.114)	(0.106)	(0.091)
IS punishment effectiveness high	-0.152	-0.158	-0.029
	(0.153)	(0.170)	(0.107)
Coexistence	0.022	0.021	0.009
	(0.154)	(0.168)	(0.561)
Six votes		-0.181	
		(0.189)	
Controls	No	No	Yes
Log-likelihood	-154.0	-182.5	-82.4
Observations	306	358	204

Note: Standard errors in brackets. Random effects probit models. Marginal effects reported. Data from robustness treatments. In regression 2, data from phases 4 and 7 of the **DC** and **DE** treatments in the main experiment are included. Phase dummies included in all regressions (not shown). Control variables are the same as those included in Table 2, regression 2. * significant at 10%; ** significant at 5%; *** significant at 1% (refer to the test of the hypothesis that the underlying coefficient is zero).