State *or* Nature? Endogenous Formal vs. Informal Sanctions in the Voluntary Provision of Public Goods[*]

Kenju Kamei, Bowling Green State University
Louis Putterman, Brown University[#]
Jean-Robert Tyran, University of Vienna and
University of Copenhagen

January 6, 2014

Abstract

We investigate the endogenous formation of sanctioning institutions supposed to improve efficiency in the voluntary provision of public goods. Our paper parallels Markussen *et al*. (forthcoming) in that our experimental subjects vote over formal vs. informal sanctions, but it goes beyond that paper by endogenizing the formal sanction scheme. We find that self-determined formal sanctions schemes are popular and efficient when they carry no up-front cost, but as in Markussen *et al*., informal sanctions are more popular and efficient than formal sanctions when adopting the latter entails such a cost. Practice improves the performance of sanction schemes: they become more targeted and deterrent with learning. Voters' characteristics, including their tendency to engage in perverse informal sanctioning, help to predict individual voting.

JEL classification codes: C92, C91, D03, D71, H41.

---

* We wish to thank the Danish research council (FSE) and the Austrian Science Fund (FWF project no. S10307-G16) for financial support. We are grateful to two anonymous referees for their detailed comments and suggestions, and we likewise thank participants at the Economic Science Association meeting in July, 2010 in Copenhagen, XVIth IEA World Congress in July, 2011 in Beijing and at workshops at Brown University and George Mason University for their helpful comments.
# Corresponding author. Department of Economics, Box B, Brown University, Providence, RI 02912. Fax: 401-863-1970; Tel.: 401-863-3837; Louis_Putterman@Brown.Edu.

Keywords: Sanction, social dilemma, public goods, voluntary contribution mechanism, punishment, experiment.

1. Introduction

Social dilemmas, in which uniformly self-interested behavior makes a social goal less rather than more attainable, are recurrent problems in modern economies. The common 'textbook' solution is the use of what we call *formal* sanctioning schemes. This solution relies on the coercive powers of the state, which can make contributions (such as tax payments) mandatory and subject to penalties for non-compliance. But formal sanction schemes are costly as they require infrastructure (like a judicial system, police and prisons) and their use seems infeasible or unnecessary in some situations. There are indeed numerous collective action dilemmas, including provision of public goods that fall beneath government notice because they are too local or parochial, for which this solution is unavailable or less efficient than what we call *informal* sanctions. Informal sanctions are decentralized and horizontal in nature and do not require formal infrastructure as they rely on peer punishment. The willingness to mete out such (costly) informal sanctions and to do so in a well-targeted way depends on the characteristics of group members, in particular their social preferences.

While the workings of formal sanctions are rather independent of social preferences once they are in place, their establishment is likely not to be. The very existence of a government that can promulgate and enforce regulations to deter free riding while acting as a faithful agent of its citizens depends on voluntary pro-social acts such as citizen scrutiny of politicians' actions, self-education about political issues, and making the effort to vote in elections. The relationship between voluntary collective action and fulfillment of obligations under threat of formal sanctions is thus a complex one, and the question of when we can or should rely on formal sanctions to resolve social dilemmas is an important and underexplored issue.

Experimental economists have extensively studied public goods problems, with a special focus on voluntary contribution dilemmas with informal sanctions (for overviews, see Gächter and Herrmann, 2008 and Chaudhuri, 2011). More centralized forms of governance, and the conditions, if any, under which informal sanctions regimes are preferable to formal ones, have received surprisingly little attention until the present paper

and a companion paper by Markussen *et al*. (2010).[1] The reason for this neglect may be that it seems obvious to an economist that appropriately designed formal sanctions dominate informal ones. We argue that which sanctions scheme is the better choice is far from obvious when agents are not fully rational and self-interested. Our results indeed show that apparently dominated informal sanctions are chosen quite often when formal sanctions are available at a moderate fixed cost, and that they are surprisingly successful if chosen. Our results also show that if voters have a say on how formal sanction schemes are to be designed, they design them to target free riders and to be deterrent in strength, which contributes to their efficiency.

In a formal sanctions (FS) regime, a group adopts a rule specifying what penalties will be imposed under what conditions and sets up a body (in large group settings, an administration or government) that observes rule violations and imposes the stipulated penalties. In an informal sanctions (IS) regime, group members can punish each other at a cost to both the punisher and the punished. Because such sanctions are costly, a rational and selfish decision-maker does not engage in IS in one-shot (and, by extension, finitely repeated) interaction, and voting for IS is therefore pointless among such agents. In contrast, voting for FS is (at least weakly) dominant for rational and self-interested agents, as long as FS target free riders reliably and the fixed cost of having FS in place does not exceed its benefit. In such a population, FS are the efficient choice.

Even when rationality, self-interest, and common knowledge assumptions are relaxed, appropriately designed FS may be preferred by voters. One reason is that in an ideal system of FS, sanctions are meted out automatically but uncertainty prevails about who will impose IS on whom, and when. The predictability of FS may mean that fewer costly sanctions need be imposed under a FS regime, an efficiency advantage. If IS do occur when allowed, there is no way to guarantee that they are not misdirected ("perverse" or "anti-social") and to rule out attempts at retaliation (see below), whereas rules in the interest of every group member can assure that FS are well targeted. This may help explain

[1] Some recent contributions on sanction systems in which a central or specialized agent is the only one empowered to sanction are discussed and compared to decentralized or informal sanction schemes by Nosenzo and Sefton (2012). In our paper, we consider only those centralized sanction schemes that are rule-based, as opposed to ones that empower a central agent to use his or her discretion. We defer discussion of "pool punishment" (Sigmund *et al*., 2010) until the next section.

why the centralized administration of penalties has traditionally been seen as the hallmark of civilization, whereas the enforcement of rules by individuals is often denigrated as "vigilante justice" or "mob rule."

There are reasons for preferring informal sanctions in some circumstances, however. We think of IS in connection with village management of woodlots and irrigation systems or social pressure in work teams, while FS bring to mind managerial structures and governments. Having an enforcing body in place—for example, a police force and courts—involves enforcement structures that, paradoxically, may need to be used relatively little if their presence suffices to deter the social bads they are meant to prevent. Informal sanctions may be less deterrent because they are less certain, but they may have the advantage of avoiding substantial fixed costs.[2]

By investigating collective choice between FS and IS, our paper contributes to a recent and rather thin experimental literature on endogenous institutions (see section 2 for references). Our design involves repeated choice between FS and IS, both when having FS in place is costly and when it is not. As a control, we also conduct sessions in which neither FS nor IS are available.

These design aspects are also present in a companion paper by Markussen, Putterman and Tyran (forthcoming). To our knowledge, these two studies constitute the first experimental research on collective choice between and performance under both FS and IS schemes. The key difference between the two studies is that whereas in Markussen *et al*., the FS scheme available to a given subject group is exogenously specified, subjects in the present paper not only choose between FS and IS but also determine the parameters of the FS, if used. This additional dimension of choice makes collective choices cognitively more demanding for voters. An important implication of this difference is that we create a more level playing field between FS and IS by allowing subjects to choose inefficient properties of the FS scheme. Voters can choose to implement either deterrent or non-deterrent sanctions, paralleling the fact that IS can be strong or weak. Voters also have

[2] To be sure, the idea that formal sanctions are more certain than informal ones does not always hold. A government might be inept, corrupt, or lacking in enforcement capacity, whereas if most members of a group were to embrace the goal in question, informal sanctioning may be a highly predictable response to norm violation.

the possibility of implementing "perverse" formal sanctions—ones that punish high contributors, similar to the "perverse" sanctions that are frequently observed in experiments with IS. Our new design also adds insight into the roles of experience and individual orientations in voting decisions.

Our study and that of Markussen *et al*. also differ in other details, including the country in which they were conducted, the level of fixed costs for having FS in place, presence or absence of variable costs of formal sanctions, and inclusion of treatments allowing exposure to each institution prior to voting. While strict comparisons are ruled out since multiple dimensions differ, each study's results can be viewed as a set of informal robustness tests of the other. The main shared results are that informal sanctions are surprisingly popular and effective, and that fixed costs of having FS in place are crucial to subjects' choices between IS and FS.

Novel findings of the present study are that subjects improve the terms of their formal sanctions schemes with learning, that they prefer deterrent over non-deterrent formal sanctions, and that subjects who exhibit cooperative tendencies are significantly more likely to vote for efficiency-enhancing institutions, and vice versa for those with uncooperative tendencies. The finding that the latter are likely to vote against efficiency-enhancing institutions, but usually as a minority, affirms a potentially important insight about heterogeneous social preferences and majority voting (Ertan *et al*., 2009). To conserve space, we leave details of Markussen *et al*.'s findings to be discussed in the context of comparisons to our own findings, below.

We proceed as follows: Section 2 discusses theoretical considerations regarding collective action and formal and informal sanctions, and briefly reviews relevant experimental research. Section 3 spells out our experimental design, the rationale behind the various treatments, and the predictions both under the assumption of common knowledge of rationality and self-interest and under alternative behavioral assumptions. Section 4 summarizes the experimental results. Section 5 concludes the paper with a brief summary and suggestions for future research.

2. Theory and literature

We consider a group with n members who engage in a finitely repeated interaction where, in each period, each individual must allocate an endowment of E money units between a private and a public (or group) account. Money placed in the public account is scaled up and divided equally among the group members, generating a marginal private return on contributions m, where $0 < m < 1$, and a return to the group of $m \cdot n$, where $1 < m \cdot n < n$. In this canonical voluntary contribution mechanism or public goods game, total earnings are maximized if each member places the full E in the public account, but it is in each individual's private interest to put zero in that account.[3]

a) Behavior under and voting on sanctions regimes, according to standard theory

Suppose that group members are asked to vote for one of two alternatives: (a) engaging in the interaction under the rules described above, and (b) engaging in an otherwise identical interaction while additionally implementing a formal sanction scheme where for each monetary unit an individual puts in her private account, she is fined s units, with $s > 1 - m$, and implementing the scheme costs each individual f per period. Then as long as $f < (m \cdot n - 1) \cdot E$ and the assumption of common knowledge of self-interest and rationality holds, group members earn more per period with than without the scheme. This is because the scheme induces a rational individual to contribute E to the public account, and with all doing this, each earns $m \cdot n \cdot E - f$ per period, which exceeds the E per period earned under option (a). An individual would accordingly choose (b) over (a) if expecting to influence the outcome with positive probability.[4]

Suppose now that there is a third option, (c), wherein the game played by the group of individuals each period has two stages. In the first stage, each individual decides how many of his or her E units to allocate to the public account. In the second stage, each individual learns the amount allocated by each of the other $n - 1$ individuals and can

[3] To be sure, models assuming infinite repetition are often applicable, because last periods can be unpredictable and reputations may carry into new settings. However, finite repetition can also be argued to characterize a variety of real world circumstances, and for an initial exploration, there are advantages to confronting the clear predictions of the finite repetition model with empirical observations in a controlled setting.

[4] It is true that there exist many subgame perfect equilibria with non-pivotal voting, under which some subjects vote for option (a). However, the possibility of errors and absence of opportunities to communicate suggests that they would reject weakly dominated choices and vote for option (b) (trembling-hand perfect equilibrium).

reduce the earnings of any of them by P units per unit deducted from her own provisional earnings for the period. Suppose that an individual i contributes C_i to the public account. Then, the individual's earnings for the period are given by

$$(E - C_i) + m \cdot \sum_{j \text{ includes } i} C_j - \sum_{\text{all } j \neq i} p_{ij} - P \cdot \sum_{\text{all } j \neq i} p_{ji} \quad (1)$$

where C_i indicates individual i's allocation to his or her public account, p_{ij} indicates the number of units of punishment that group member i gives to member j and conversely for p_{ji}, so that the third term is i's expenditure on punishing others and the fourth is i's losses from being punished by others. In this case, if there is a threat of receiving more than (1 – m) units of punishment for each unit of endowment placed in one's private account, it is privately optimal to contribute everything to the public account. However, in a finitely repeated game with rational, strictly self-interested agents having common knowledge that all are of the same type, there can be no such credible threat, since it is never privately optimal to punish in the last period and therefore punishing in any earlier period as a way of threatening to punish future free riding is not credible. In this setting, the availability of a punishment option makes no difference, and uniform free riding with the associated earnings of E per period prevails. Assuming uniform voting for the weakly dominant alternative, then, standard theory predicts that option (b) will be selected in favor of either of options (c) and (a), while individuals will be indifferent between options (a) and (c). (In our experiment, subjects will vote only between (b) and (c), so the operational prediction is that they will vote for (b).)

b) Possible behavioral departures

Observations from past public goods experiments suggest that standard theory does poorly in predicting many decisions in settings of types (a) (simple VCM) and (c) (VCM with punishment stage). Typically, subjects are found to contribute 40 to 60% of their endowments to the public account on average in the initial period (Zelmer, 2003). Contributions trend downwards with repetition, but are rarely uniformly zero even in the known last period. There is evidence of a "conditional willingness to cooperate" on the parts of many subjects (e.g., Fischbacher and Gächter, 2010 and Thöni *et al*., 2012). In experiments of type (c), most subjects give costly punishment at least once over the course

of repeated play (Falk *et al*., 2005 and sources cited above). While subjects mainly direct punishment at the lower contributors in their groups, many subject pools also exhibit substantial punishment of high contributors, which Cinyabuguma *et al*. (2006) dub "perverse punishment" (see also Ertan *et al*., 2009).[5] Contrary to the prediction above that the availability of punishment does not alter equilibrium play, its availability at sufficiently low cost is associated with at least sustained if not rising contributions to the public account in numerous experiments (see, e.g., Nikiforakis and Normann, 2008).

While less studied, the impact on contributions of formal sanctions that make contributing privately optimal appears largely consistent with the standard theoretical prediction (Putterman *et al*., 2011). The main departure from the standard predictions regarding formal sanctions is that subjects significantly increase their contributions in response to sanctions of too small a magnitude to render contributing privately profitable—i.e., "non-deterrent sanctions"—when these have been selected by the subjects themselves (Tyran and Feld, 2006).

Two final observations that may be pertinent to predicting behavior in our experiment revolve around voting. First, the selection of an institution by voting may in and of itself influence its performance (Tyran and Feld, 2006, Dal Bó *et al*., 2010, Kamei, 2013, Sutter *et al*., 2010). This factor might also apply to informal sanctions. For example, a group vote to make informal sanctions available might be interpreted as a signal that most members are willing to punish free-riding and that they look favorably on others doing so, which might cause increased deterrence for given sanctions as well as better targeting of whatever sanctions are imposed. A favorable impact of voting on performance of IS would affect our expectations about institutional choice because if efficiency is high under IS, for which no fixed cost is paid, the cost that a rational subject should be willing to pay to have a formal sanction scheme in place is less than the initially predicted $(m \cdot n - 1) \cdot E$.

[5] Herrmann *et al*. (2008) refer to the overlapping phenomenon of a group member punishing one whose contribution was higher than her own as "anti-social punishment."

Second, predicting the impact of voting is somewhat more complicated in our experiment than in others, including Markussen *et al*., because use of the formal sanction option by our subjects always entails first a vote on whether to employ such a scheme and then a vote on its precise terms. Since a vote for the formal scheme means entrusting a majority of fellow group members with the power to set parameters that will be binding on oneself, beliefs about others' intentions and rationality assume considerable importance. If we find frequent choice of IS in our design, it could thus in part be reflecting subjects' uncertainties about one another's next stage voting behaviors.

Finally, formal social preference models can be used to generate predictions of behavior by subjects of varying type. Markussen *et al*. do this using a modified version of a model in Charness and Rabin (2002) (and in their appendix, also Fehr and Schmidt, 1999). Many of their predictions carry over to our setting and measure up well against both sets of data; we refer the reader to their paper for details.

c) Literature

Experimental studies of endogenous choice of institutions for collective action began with work of Yamagishi (1986), who offered subjects in a version of the public goods game in which the return from own contribution is zero the opportunity to voluntarily fund a mechanism for penalizing free riders. Any money placed in the punishment fund generated twice as large a loss for the lowest contributor to the public good. The mechanism proved successful in raising contributions to the public good. Another important study is Ostrom *et al*. (1992), who let subjects communicate and decide whether to use sanctions on overharvesting to achieve greater efficiency in a common pool resource game. The combination of communication and sanctions is effective at raising efficiency, in their study.

More recently, Botelho *et al*. (2005), Gürerk *et al*. (2006), Ertan *et al*. (2009), and Sutter *et al*. (2010) have studied whether subjects will choose to face a voluntary contribution dilemma that permits or one that does not permit informal sanctions. Botelho *et al*. and Sutter *et al*. implement experimental designs in which subjects make a single choice over using or not using an informal sanction mechanism (and, in Sutter *et al*. also informal rewards). Both papers find informal sanctions to be unpopular. Gürerk *et al*. and

Ertan *et al*.'s designs include a series of choices between sanction-free and sanction-permitting schemes, and both find that informal sanctions are unpopular at the outset but become increasingly popular with experience—which indicates that evolution of institutional choices may also be important to study. Our study's institutional choice mechanism, voting, makes it more like Ertan *et al*., e.g., than like Gürerk *et al*., which uses 'voting with one's feet,' and Yamagishi, in which institutional choice is associated with the voluntary funding of an exogenously present mechanism for which no alternatives are offered.

A more recent study by Traulsen *et al*. (2012) builds on Yamagishi's by comparing a formal style punishment mechanism, "pool punishment," to an informal or individual punishment opportunity, adding punishment of those who fail to punish in some treatments. In one treatment, subjects individually select between the two kinds of punishment, and pool punishment emerges most popular. Traulsen *et al*. and the closely related theoretical study of Sigmund *et al*. (2010) resemble our research in that they study choice of up-front and centralized vs. *ex post* and decentralized sanctioning, but they differ in that contributions to the centralized punishment pool remain voluntary, that the institutional choice emerges without voting, and that the targeting and intensity of pool punishment are exogenous.[6] Traulsen *et al*. report that their "experimental results show how organized punishment could have displaced individual punishment in human societies."

The study of institutional choice through voting on formal sanctions begins with Tyran and Feld (2006), who let some subjects vote on whether to impose on themselves a "mild law"—i.e., non-deterrent sanction. In another treatment, the same sanction is imposed or not imposed exogenously. Some 60% of groups offered the choice choose the mild law and achieve higher efficiency than either those in the exogenous or endogenous no law and in the exogenous mild law conditions. In an experiment on spillover effects of democracy, Kamei (2013) also lets subjects vote on the introduction of a mild sanction in a VCM, finding that pro-sanction subjects cooperated significantly more when the law was endogenously, rather than exogenously, imposed.

[6] Also closely related is Zhang *et al*. (forthcoming), which we discuss it briefly in footnote 26, below.

To our knowledge, studying choice by vote between informal and formal sanction regimes in the VCM commences with Markussen *et al*. (forthcoming) and the present paper. Subjects in the former decide which institution to use (and in some treatments, also consider the option of a sanction free institution), but the terms of the formal sanction regime are dictated exogenously, with between-subject variation. While hypotheses on voting are more easily derived by formal analysis in such an environment, studying choice of institutions when the sanctions schemes in question are structured by those using them is nonetheless of interest in its own right. With endogenous choice of scheme parameters, as mentioned, subjects can vote to sanction high contributors, just as some individuals punish perversely in an informal sanctions regime. We can also obtain observations of individuals interacting under an informal sanctions regime and test whether tendencies to cooperate by contributing or to behave anti-socially by punishing high contributors predict votes about the formal scheme's parameters. We can examine, too, choices between deterrent and non-deterrent sanctions and investigate how decisions about scheme parameters change over time.

Studying choice between formal and informal sanction schemes when formal scheme parameters are endogenous is also important as a check on a potential bias that the formal schemes used in Markussen *et al*. might introduce. Because the available formal schemes there specify punishment of low contributors only, the very presence of this option and its explication to subjects in the instructions might have cued subjects to the idea that punishing low contributors is what sanctions (including informal ones) are for. Most experimental instructions used by economists since Fehr and Gächter (2000), by contrast, have worked hard to avoid any such suggestion in order that any sanctioning that emerges be truly endogenous. By offering subjects more general-purpose formal schemes that can be used to penalize either high or low contributors, our design preserves the more neutral environment of the earlier cooperation-and-punishment literature. This allows us to check whether the generally efficient performance of informal sanctions in Markussen *et al*. was substantially influenced by the restricted nature of the formal schemes available.

A first step towards studying the endogenous design of a formal sanction scheme is taken in Putterman *et al*. (2011). There, presence of an exogenous formal sanctions scheme is given, but it is left to subjects to determine by vote the level of sanctions and whether it

is contributions to the public or to the private accounts that are subject to them. In the event, seven of eight groups select fully efficient sanction schemes by their third of five votes, indicating that most subjects gravitate towards efficient sanctions but that some learning is entailed.[7]

3. Experimental Design

At the heart of our experiment are six VCM treatments in which subjects have opportunities to choose between informal and formal sanction schemes. In two of these treatments participants are inexperienced with the sanction mechanisms when voting for the first time, whereas in the other four, participants are exogenously assigned substantial experience with each scheme before voting, affording a check of whether experienced subjects vote differently from inexperienced ones.[8] The experiment also includes a main control treatment without sanctions, and three special control treatments discussed later.

In all treatments, groups have $n = 5$ subjects (partner matching). Each subject is provided at the beginning of each period with a fixed endowment of $E = 20$ points (34 points = \$1), and is asked to make a series of decisions on the allocation of that endowment between a private and a public account. The marginal per capita return (MPCR) is $m = 0.4$. We set the number of periods of play at 24 to allow for the evolution of institutional choice over time, and we group the periods into six phases of four periods, with a group's institution remaining fixed within a phase. Given this structure, subjects choose institution by voting 3 times in the treatments in which they are first exposed to a no-sanctions regime (NS), FS and IS exogenously (3-Vote treatments), but do so 6 times in treatments without such exposure (6-Vote treatments). Voting is simultaneous, mandatory, and free, with each subject indicating a preference for either IS or FS and groups learning the outcome of the majority vote but not the number of subjects voting for it.

[7] Because of the greater overall complexity of our design, we simplified choices regarding the formal scheme relative to those in Putterman *et al*., reducing the number of decisions required from three to two. In particular, subjects in that paper but not the present one are asked to determine the scope of a potentially penalty free range ("exemption"), whereas this possibility is not mentioned and hence the exemption range is automatically of size zero in our design.

[8] For example, subjects in Gürerk *et al*., Ertan *et al*., and Markussen *et al*. show initial reluctance to adopt IS. If IS is beneficial to most subjects in practice, then by having subjects experience IS early on exogenously, the 3-vote treatments might induce a higher proportion of votes for it than do the 6-vote ones, with voting from the start.

Table 1 provides an overview of our six main and three additional treatments, with the rows distinguishing number of votes and order of exogenous conditions, the columns, whether there is a fixed cost of adopting a formal sanction scheme. In all treatments, subjects start by playing a VCM in an NS regime for at least one period so as to familiarize them with the nature of the social dilemma. In 3-Vote (see upper panels), subjects play a series of three 4 period phases, the first under exogenously imposed NS, followed by either a phase under IS and one under FS (dubbed IF order, upper two cells) or a phase under FS and one under IS (next two cells, dubbed FI order). In the last three phases, groups play with either FS or IS, depending on the outcome of the vote in their group. The FS treatments in the left column carry no fixed administrative cost, those in the right column a fixed cost of 5 points per group member and period. The treatment order and administrative cost (none = N, 5 point cost = C) is indicated as follows: 3(IF)-N, 3(IF)-C, 3(FI)-N and 3(FI)-C.[9]

In the 6-Vote treatments, subjects play a single period in NS condition followed by six phases, played under each group's majority choice of either FS or IS, also determined in start-of-phase votes. Since there are no exogenous order differences among the 6-Vote treatments, they are distinguished only as 6-N and 6-C. The sanction-free BASELINE treatment and the three additional treatments listed in the dashed cells of Table 1 are discussed later.

Figure 1 shows the timing of events.[10] When IS are in place, the contribution stage is followed by a punishment stage in which each subject i assigns points $p_{ij} \in \{0, 1, 2, \ldots, 10\}$ to each other group member j. The contribution of each j is shown in random order. Each point of punishment costs the recipient four points ($P = 4$ in Equation (1)), except if received punishment exceeds first stage earnings, in which case first stage earnings minus

[9] While group size, endowments, and MPCR are identical in our study and Markussen *et al*., our fixed costs of 0 and 5 differ from theirs, 2 and 8, allowing in an approximate sense for a kind of meta-robustness check. In the event, both studies find similar differences between their low (0 or 2) and their high (5 or 8) cost treatments, while between the two studies, the share of cooperative surplus needing to be spent by each subject in order to adopt FS includes 0% (cost of 0), 10% (2), 25% (5) and 40% (8). Unlike our design, however, subjects in Markussen *et al*. face no variable cost of FS (see below).

[10] In all treatments, subjects are informed at the outset about the number of phases and periods and that they will interact in the same group of five anonymous participants throughout. In 3-Vote treatments, subjects learn about the conditions in four distinct sets of instructions, one before each phase of exogenous play and one before the start of voting between schemes, whereas in 6-Vote treatments, they receive one set of instructions before the initial phase and a second set of instructions explaining all of the remaining elements before the second phase. See the Appendix for further details.

received punishment is set to 0.[11] Period earnings could nevertheless become negative because the cost of imposing punishment is always incurred, with any one period's losses being taken from the accumulated earnings of other periods. Thus, period earnings under IS are given by

$$\max\left\{(20 - C_i) + 0.4 \cdot \sum_{j \text{ includes } i} C_j - 4 \cdot \sum_{\text{all } j \neq i} p_{ij}, 0\right\} - \sum_{\text{all } j \neq i} p_{ji} \qquad (2)$$

which differs from Eq. (1) by substitution of specific values for *E*, *m* and *P* and by the bound on the cost of punishment received, as indicated by the max function. [12]

Having selected the FS regime, groups determine what action is sanctioned and at what rate. In each FS period, subjects first vote on whether to sanction allocations to the private or the public account and learn the outcome of the majority vote. Each then chooses a sanction rate from the set 0, 0.4, 0.8 and 1.2 points, with the median rate being implemented for (and it alone revealed to) the group. Subjects then make their contribution decisions, are shown one another's contributions in a random order, and learn their earnings for the period. For each point a group member allocates to the account in question, he loses an amount equal to the chosen sanction rate with certainty. Note that when contributions to the private account are sanctioned, a subject's privately optimal strategy remains zero contribution if a rate of 0 or 0.4 is set but is contributing her full endowment if the rate is set at 0.8 or 1.2 points (see equation (3)).

One reason Markussen *et al*. made the terms of FS regimes exogenous was to avoid raising the possibility of strategic voting, which complicates theoretical predictions. If all subjects are strictly self-interested, perfectly rational, and have common knowledge of this, and if we adopt the assumption that each subject votes her preferences due to the

[11] See Nikiforakis and Normann (2008) on the relation between efficiency and the cost ratio *P*, including at a 1:4 ratio. This ratio is also used in numerous studies, including Page *et al*. (2005), Bochet *et al.* (2006), Önes and Putterman (2007), and Sutter *et al*. (2010). An effective ratio of 1:4 or more often obtains for the first point of punishment *i* assigns to *j* in Fehr and Gächter (2000) and in the numerous experiments adopting its punishment schedule. Markussen *et al*. (forthcoming) mainly use the same 1:4 ratio but also consider robustness to lower sanction effectiveness, finding surprisingly little effect.

[12] In practice, maximum punishment is rare and nullifies first-stage incomes only occasionally. For example, of the 1090 instances in which a subject *i* punished another subject *j*, only 5 (0.5%) involved giving 10 points of punishment. Punishment received exceeded first stage earnings in only 23 of 2960 subject-periods (0.8%). Finally, punishment almost never resulted in losses to the punisher. Only 5 of the 2960 subject-periods (0.2%) under IS saw a subject incur negative earnings for the period, and all subjects had strictly positive earnings overall.

possibility of being pivotal, then strategic voting should not be an issue, because it is the best outcome for each that there be a deterrent formal sanction, which shifts the predicted equilibrium from full free riding and earning 20 points per period to full contributions and earning 40 minus the fixed cost (0 or 5) per period. However, dropping any of the three assumptions (self-interest, rationality, common knowledge) makes strategic voting a possible concern. For example, subjects who for some reason want to avoid having any sanctions at all may vote for FS, then if the group has an FS majority outcome, vote for penalizing high contributors, in the expectation that if that choice prevails, others will join in selecting a sanction rate of zero. We undertook our experimental study despite the difficulty of providing exhaustive voting predictions because we believe that studying institutional choice in this somewhat more realistic setting is an important empirical complement to Markussen *et al*.'s approach.

For parallelism with IS, we made groups bear a variable cost of imposing sanctions, as well as the possible fixed cost. If a group member is sanctioned, one third of the amount deducted from that member's earnings is also charged to the group as a whole and is divided equally among the five members, meaning each individually pays an amount equaling (1/3)*(1/5)=1/15 of the sanctioned individual's loss. With the other four collectively paying 4/15 of the sanction loss and the sanctioned member bearing the same cost of sanction imposition, for a total loss of (16/15) times the sanction, the ratio of cost collectively born by the others to cost to the sanctioned individual is thus 1:4, exactly as in IS.[13] The cost of imposing sanctions, including the fixed cost, is always borne by each individual even if it makes her earnings for the period negative, whereas the effective sanction itself cannot exceed the individual's first stage earnings for the period—both features being as in IS. Thus, an individual's earnings under FS are given by

$$\max\left\{(20-C_i)+0.4\cdot\sum_{j \text{ includes } i} C_j \; - p_i, 0\right\} - \frac{1}{15}\sum_{j \text{ includes } i} p_j - f \qquad (3)$$

[13] If the fixed costs of FS can be thought of as representing among other things, costs of building and maintaining prisons and court houses, the variable ones might correspond to the costs of pursuing criminals and conducting trials, which depend on how many rule violations occur. We made the ratio the same for both kinds of sanctions and referred to the uniformity of 1:4 ratio under the two schemes in the instructions so that the difference in variable cost per sanction would not in itself influence subjects' votes (see Appendix A).

where $p_i = r \cdot (20 - C_i)$ is the imposed sanction if contributing to the private account is penalized,[14] with $r \in \{0.0, 0.4, 0.8, 1.2\}$ being the sanction rate, and $f \in \{0, 5\}$, depending on the treatment, the fixed or administrative cost of FS. As should be clear from equations (2) and (3), the costs of a sanction scheme and of any fines paid are lost and are not redistributed to subjects in any way.

In the BASELINE treatment, subjects play the standard VCM game in NS mode for all six phases. To control for restart effects (Andreoni, 1988), phases are separated by a break of 40 seconds to parallel the breaks for voting or instructions in the other treatments. As in the other treatments, feedback about other individuals' contributions to the public account is presented in a random order each period.

We try to present instructions for all treatments in natural language but avoid terms that might suggest that contributing to the public account is desirable. For example, we use "allocate" rather than "contribute," "assign reduction points" rather than "punish" or "sanction." In an exception to our verbal neutrality, we use "fine" to describe the FS scheme, but we indicate that this can be a fine for allocating points to the private or to the public account. The schemes voted on are called "individual reduction decisions" (= IS) and "group-determined fines" (= FS), and the fixed cost of FS, in the treatments including it, is called a "fixed administrative cost of having a fine scheme in operation."

All treatments also include one additional task before the main experiment and two at the end. These tasks were included to gauge subjects' conditional inclinations to cooperate, IQ, and political orientation, factors which help to explain the heterogeneity of behavior in some settings. [15]

4. Results

The experiment

Twenty sessions, including six for the additional treatments discussed later, were conducted in a computer lab at Brown University during 2009 -2011 with a total of 375

[14] $p_i = r \cdot C_i$ if contributing to the public account is penalized.

[15] Our online Appendix provides the full instructions for these tasks. The indicators constructed from them are as in Putterman *et al*. (2011). An exception to the statement that all subjects completed these tasks concerns subjects in the Fuller Information treatments discussed at the end of Section 4. The IQ and political orientation questions were left out of those treatments to conserve time.

undergraduate subjects drawn from the full range of disciplines.[16] Subjects were paid privately in cash at the end of their session, with average earnings of $24.50 for the main portion of the experiment.[17]

a) Voting for formal vs. informal sanctions

Summary: Three findings on this institutional choice stand out. First, formal sanctions are not as popular as predicted by standard economic theory. Second, fixed cost matters more than such theory predicts. Third, the order in which schemes are exogenously experienced has no significant effect.

Figure 2(a) shows the pattern of voting outcomes over time, with the data grouped into 3-Vote and 6-Vote treatments and treatments with and without administrative cost in the FS scheme. A summary of individual votes and group outcomes is provided in Appendix Table B.2. Contrary to the prediction of standard theory that subjects would uniformly select formal over informal sanctions, the choice between the two schemes favors formal sanctions over informal ones by a relatively narrow margin (102 group votes versus 81 group votes), with the administrative cost of FS serving as a major discouragement of that system's use. Whereas 75 out of 87 group votes (86%) favored FS over IS in the no administrative cost treatments, the corresponding numbers were 27 out of 96 votes (28%) in treatments with 5 point administrative cost to operate FS. The relative popularities of the two institutions and the impact of the fixed cost of FS parallel the findings of Markussen *et al*., demonstrating the applicability of their findings that IS is popular and that fixed cost is the key influence on institutional choice over a broader range of fixed costs and, more importantly, to environments in which subjects control the terms of the formal sanction scheme.

The order in which subjects are exogenously exposed to the FS and IS conditions in the 3-Vote treatments appears to have had no effect on choice of scheme in phases 4, 5

[16] An experimenter read all instructions aloud as participants read along. Subjects answered control questions after each set of instructions to test comprehension. In 3-Vote treatments, questions were taken and answered after the reading of instructions before each of phases 1 though 4. In 6-Vote treatments, questions were taken and answered after the reading of instructions before phases 0 and 1, and in BASELINE, after the reading of the only instructions, those before phase 1. The experiment was programmed in z-Tree (Fischbacher, 2007).

[17] All subjects also earned a show-up fee of $5.00. In the IQ portion completed by the 300 subjects in the main treatments, earnings averaged $2.20.

and 6 (see Appendix B.2). We therefore pool the two 3-N treatments and the two 3-C treatments in the remainder of our analysis, although an order dummy variable is included when we estimate regressions.

b) Voting on what to sanction in the formal scheme

Summary: Although only a few voting outcomes are for sanctioning allocations to the public account, about an eighth of individual votes in total are cast for this option. In treatments in which exogenous punishing opportunities precede voting (3-C, 3-N), perverse informal punishers vote this way significantly more often than do other subjects. When sanctioning allocations to the private accounts, groups usually pick the most severe sanction available.

The 102 group votes in favor of FS are each followed by four period-specific votes on that scheme's two parameters, for a total of 408 group and 2,040 individual votes on each dimension. There are another 124 group and 620 individual votes on each parameter during periods of exogenous FS play in the 3-vote treatments. 97.7% of group voting outcomes are for sanctioning contributions to the private rather than public accounts (see Appendix Table B.1). Still, 13.3% of individual votes favor sanctioning contributions to the public accounts—a proportion of "perverse voting" that approaches but is somewhat less than the share of punishment perversely directed at high contributors in past experiments with the same university subject pool (e.g. Ertan *et al*.).

We estimate random effects probit regressions to investigate what individual-specific factors account for subjects' votes to penalize contributions to the public versus the private accounts, when FS is in place. Estimating separately by treatment and for periods of endogenous versus exogenous FS, our explanatory variables are subjects' average conditional contribution, IQ, political orientation, gender, an indicator for engaging in perverse punishment and another for having received such punishment during the exogenous IS phase, and period. In all treatments for which the perverse punishment variables are obtained, perverse punishers are significantly more likely to vote to penalize contributions to the public rather than private accounts (in other words, to vote for perverse formal sanctions). In the regressions for the 3-N treatment, higher average conditional contribution is associated with voting to penalize contributions to the private

accounts (i.e., "voting non-perversely"), significant at the 5% or 10% levels. (For details, see Appendix Table B.3). The remaining variables are usually insignificant. Thus, we find evidence that individuals' orientations towards cooperation, as indicated by both contribution and punishment behavior, impact their voting on the targeting of the formal sanction scheme. The out-voting of subjects disinclined to cooperate by ones favorably inclined towards cooperation helps to explain why almost 98% of majority votes supported efficient targeting.

c) Voting on the formal sanction rate

Following the 520 group votes for penalizing allocations to the private accounts, about 90 percent of groups chose a binding sanction (433 chose penalty rate 1.2 and 34 chose rate 0.8).[18] In contrast, in every period in which a group had chosen to penalize contributions to their public account, its majority set the penalty rate at 0.

Figure 2(b) shows the proportion of median votes cast, thus the rates that went into effect, among the observed options. One interesting observation is that while choosing to penalize contributions to private accounts at the maximum rate of 1.2 is the preponderant choice (89.2% of votes) in the treatments without administrative cost, its dominance is less overwhelming (65.1% of votes) in the treatments with administrative cost, perhaps because subjects were reluctant to add to the fixed cost of 5 points per period that they already bore when using FS (a kind of sunk cost fallacy, if $r < 0.8$ is chosen). There are also indications of a learning process, with the share of (FS-using) groups fining contributions to private accounts at the maximum rate rising from less than half in Phase 1 to 100% in Phases 5 and 6 of the 6-vote treatments.[19]

d) Contribution levels and trends by condition

[18] At the individual level, in cases in which formal sanctions were directed at contributing to the private accounts, 71.3% of votes were for the 1.2 rate, 5.4% for 0.8, 4.5% for 0.4 and 18.9 % for 0.0.

[19] See Appendix Table B.2 (d). In our working paper, we discuss regressions using individual characteristics to predict subjects' votes on the sanction rate when contributions to the private accounts are fined; see Appendix Table B.4. Estimates in several of the regressions suggest that either higher IQ or more conditionally cooperative subjects or both tended to vote for higher sanction rates. While the number of periods and groups in which contributions to the public account were sanctioned was too small to analyze using regressions, we note that only 15% of the relevant votes on rates were for positive sanction rates, and that those votes came disproportionately from participants who were perverse punishers during play under IS.

Summary: Both IS and FS increase contributions, with the two being equally effective in later phases. FS also increase contributions in cases in which a theoretically non-deterrent sanction rate is chosen.

The left panel of Figure 3 displays the average contribution to the public account by period and condition, distinguishing between formal and informal sanctions schemes when imposed exogenously (in Phases 2 and 3 of 3-Vote treatments) versus when adopted by vote (all phases of 6-Vote treatments and remaining phases of 3-Vote treatments). The dashed curve shows that in the BASELINE treatment and in Phase 1 of the 3-Vote treatments, with no sanctions (NS), contributions follow the familiar pattern of beginning around 50% of endowment and declining with repetition.[20] The remaining curves, in contrast, show that groups interacting under either sanctions regime, whether with exogenous or endogenous regime choice, exhibit strong upward trends in contributions in Phases 1 – 3 and stable and high contributions above 90% of endowment in Phases 4 – 6. In all phases providing sufficient observations for testing, contributions are significantly higher with either IS or FS than in both Phase 1 and BASELINE treatment NS.

Comparing the impacts of IS and FS, we find that with both voted and exogenous sanctions, average contributions are higher in the first three phases of using sanctions when FS is used than when IS is used. This difference is sometimes statistically significant, but always economically small relative to the difference between either regime and NS (see Figure 3 (a)). In the last three phases, moreover, contributions do not significantly differ between FS and IS (see Appendix for details).

To see how the severity of sanctions affects contributions, consider Figure 4, which shows average contribution by phase under formal sanctions, disaggregated by sanctions rate but pooling observations from all treatments (including those in which formal sanctions are imposed exogenously). For reference, average contribution under NS (in Phase 1 of 3-Vote treatments and all phases of BASELINE treatment) is indicated by the dots in the middle of each set of bars (with the dashed line indicating the contribution

[20] The fact that contributions fluctuate without trend during periods 5 – 18 is also not unusual given the partner matching protocol in which there may be attempts to restart cooperation, possibly facilitated by the pause in play following each set of four periods (see again Andreoni, 1988).

trend). Although the average contribution is lower at sanction rate 0.4 than at the higher, formally deterrent rates, the difference in early phases is surprisingly small. In all phases in which they are observed, FS of rate 0.4 yields higher contributions than NS, a finding consistent with past evidence of sensitivity of contributions to the MPCR but that may also reflect a signalling or voting effect, as with the non-deterrent sanctions in Tyran and Feld (2006). Average contributions are also higher at sanction rate 0 of the FS regime than in the NS regime in five out of six phases, but the difference is generally small, and there are too few group level observations to test for significance.

e) Punishment under informal sanctions condition

Summary: Punishment in IS is mostly targeted at low contributors and is effective in that those who are punished increase their contributions.

75.3% of subjects who play at least one phase in IS condition engage in costly punishment at least once, with 53.4% (17.4%) engaging in costly punishment in more than one out of four (one out of two) of those occasions in which group members contributed unequal amounts. Panel (b) of Figure 3 shows, separately for endogenous and exogenous IS, the average informal sanctions given per subject (one fourth the amount lost by those targeted). For both endogenous and exogenous IS, sanctions begin at around 1.6 points per subject and decline with repetition, reaching 0.4 or less points per subject in later periods of endogenous IS. The amount of perverse punishment, also displayed, is relatively small: on average, less than 8% of punishment points goes to groups' above-median contributors. Regressions (for details, see Appendix Table B.6) show punishment of low contributors in the previous period to be strongly associated with increases in their contributions, while punishment of high contributors is associated with mild decreases, justifying its description as perverse. The uptick in punishment in period 24 confirms that punishment is not entirely strategic in nature (compare Falk *et al*., 2005).

f) Effect of sanctions on earnings

Summary: Practice with IS improve performance. In later phases, IS generate higher earnings than FS with fixed costs (and about the same as FS without fixed costs). IS are in fact so well-targeted that they make contributing profitable.

In the social dilemma faced by our subjects, each can earn 40 points per period if all contribute their full endowments to the public accounts versus 20 if all fully free ride, as is individually rational taking one another's contributions as given. The possibility of sanctions can resolve this dilemma, but does so most efficiently when sanctioning costs are avoided, i.e. when the expectation of sanctions suffices to deter free-riding.

Figure 5 shows average earnings in early and late phases of the experiment under FS, IS, and NS conditions, with panels (a) and (b) showing the data for 3-Vote treatments respectively without and with administrative cost, and panels (c) and (d) the data for 6-Vote treatments likewise distinguished. For purposes of comparison, average earnings in the BASELINE treatment are also shown (identically) in each panel. We group the early and late phases to check for impacts of experience or learning. The figure shows that the earnings achieved under IS were appreciably higher in the later than in the earlier phases in both 3- and 6-vote treatments, while earnings under FS show smaller gains, mainly in the 6-Vote treatments. Wilcoxon signed rank tests find that the change in earnings over time is statistically significant for the IS but not the FS condition.

Comparing IS and FS earnings during given phases, the figure shows higher average earnings under FS than under IS in the early periods.[21] In phases 4 – 6, average earnings under IS exceed those with FS in treatments with administrative cost, with the difference being statistically significant in the 3-vote treatments. Even when there is no administrative cost, IS achieves similar efficiency to FS after experience, although it is not chosen often enough to test for statistical significance of the difference.

An observed feature of IS experiments is that efficiency may increase over time if contributions rise while the amount of sanctioning falls (see Gächter, Renner and Sefton, 2008). Such a pattern is exhibited by the data in Figure 3, and its impact on earnings is confirmed by Figure 5. On average, subjects earned 31.1 points per period during the first

[21] The difference is significant in Phase 2 and also in Phase 3 for the 3-Vote treatments without administrative cost, but not for those with that cost. The test can't be carried out for the **6-N** treatment because only one group voted for IS. See Appendix Table B.8.

phase in which their group used IS, versus 36.6 in the last such phase, a difference statistically significant at the 0.05 level according to a Wilcoxon Signed Rank test.[22]

An interesting question regarding informal sanctions is whether earnings are on average higher for those members of a group who contribute more, due to the presence of punishment. To investigate, we estimate subject-level regressions with individual fixed effects and robust standard errors clustered by group. The dependent variable is earnings and the independent variables are individual contribution and a constant. We find (see Appendix Table B.9, part (1)) that contribution is a significant positive predictor of earnings, meaning that informal sanctions did in fact reverse the incentive to free ride.

g) Explaining individuals' votes between institutions

Summary: Following exogenous experience, subjects tend to vote for the scheme under which they had earned more on average, without significant impact of earnings variation or of individual-specific characteristics. An exception is that higher IQ subjects are more likely to vote for IS in the C and for FS in the N treatments, correctly anticipating the scheme that yields higher observed earnings under the relevant fixed cost.

Figure 5 shows that in the second half of each session, whether earnings were higher under IS or FS depends largely on the absence or presence of an administrative cost when FS is used. The preponderance of voting for IS in the C and for FS in the N treatments looks rational in view of these patterns. Because neither sanction scheme dominates the other at all times and because subjects received no feedback about outcomes in groups other than their own, it remains interesting to study the impact of earnings on voting between schemes by estimating a series of probit regressions that also control for possible effects of cooperative orientation, intelligence, and other individual characteristics.

[22] The calculation and test includes any group using IS in at least two phases, regardless of whether exogenous or endogenous. We also tested for a trend in earnings under IS by estimating a linear regression whose dependent variable is average earnings per group per phase using IS, and whose independent variable is the number of phases the group has used IS thus far {=1, 2, …}. We include group fixed effects. The number of phases the group has used IS obtains a positive coefficient of 1.75 which is significant at the 1% level.

The regressions, shown in Appendix Table B.10, use observations for the votes of Phases 4 – 6 of the 3-vote treatments, in which a measure of relative earnings under both schemes is always available thanks to the early exogenous phases. We pool observations across exogenous condition orders while including a dummy for order, but we separate the observations of treatments without and with administrative cost. Except in our most basic specification, we include the ratio of the coefficients of variation of earnings under the two schemes, as experienced by the individual subject, to check whether variability of earnings matters. Additional regressions add as controls binary indicators of whether a subject gave or received perverse punishment during exogenously imposed periods under IS, with or without controls for three personal characteristics. Self-reported political orientation is included to see whether more politically conservative individuals are less supportive of the intervention entailed by FS, as in Putterman *et al*. (2011). The other included characteristics are IQ and gender.[23]

All estimates have significant positive coefficients on past relative earnings, indicating that subjects indeed tended to vote for the system under which they had experienced higher earnings on average. Coefficients on the treatment order dummy and on the ratio of experienced coefficients of variation are uniformly insignificant. In both regressions for the treatments without FS fixed cost, subjects who displayed anti-social inclinations by punishing high contributors are found to be significantly less likely to vote for a formal sanction scheme. More interestingly, all four coefficients on IQ are significant at the 5% level or better, but with opposite signs in the fixed cost and no fixed cost treatments. Higher IQ subjects are thus more likely to favor the scheme that proves most efficient *ex post* in each treatment. None of the other variables is statistically significant.

h) Did endogenous choice increase the efficiency of sanctions?

Summary: Endogenous IS are more effective than imposed IS, although the difference is only marginally significant. The result may be due in part to self-selection of cooperative types into IS.

[23] Average conditional contribution is simply the average of the individual's 21 entries in the form indicating what he or she chooses to contribute assuming others on average contribute 0, 1, …, 20. While Fischbacher *et al*. (2001) classify subjects as conditional cooperators, free riders, etc., in Putterman *et al*. (2011) we find average conditional contribution to be as good an indicator of conditional willingness to cooperate as individual type dummies or other measures based on the conditional contribution schedule.

Because the levels of contributions and efficiency observed under endogenous IS regimes in the 6-Vote treatments and in phases 4 – 6 of the 3-Vote treatments are relatively high compared to others in the literature, it is possible that part of the efficiency of voted IS could be due to a signaling or other effect of having been chosen democratically. To investigate, we conducted two sessions, each with three groups, collecting six observations of a treatment in which groups operated under the IS condition as a result of its exogenous imposition, rather than of voting. The test treatment (the Exogenous IS Comparison Treatment) was structured identically to treatment 6-C—which provides the most play under voted IS— except that in the instructions before Phase 1, when the IS and FS conditions were explained to subjects, they were told that the condition under which they would play each phase would be determined by the computer, with no mention being made of the possibility of voting.[24] In fact, the computer program assigned the subjects in all groups of this treatment to the IS condition in every phase, so that six group-level observations could be collected paralleling the groups in the 6-C treatment that voted for, and thus played under, the IS condition in each of phases 1– 6.

Appendix Figure B.4 plots average contribution and earnings by period in the four 6-C groups that voted to use IS in all periods and in the six groups of the Exogenous IS Comparison Treatment. On average, both contributions and earnings are considerably higher in 6-C groups, which used IS by choice, than in the comparison treatment groups that did so exogenously. The difference is considerable in economic terms—more than 6 points (30% of endowment) on average (see Appendix Figure B.4). Due to performance variation among the observations of each treatment, however, the differences are statistically significant only at the 10% level and only if one-tailed tests are used, as would be appropriate for testing a one-sided hypothesis that endogenous choice raises contributions and earnings.[25] Interestingly, we find indications that more cooperatively

[24] Subjects in the comparison treatment had not participated in any other treatment and with high likelihood had no knowledge that similar experiments had been conducted that included voting. The instructions described both IS and FS schemes and indicated that one or the other would be assigned in any given phase, with the assignment not being influenced by subjects' behaviors.

[25] We report Mann-Whitney tests at group level in Appendix Table B.12. We test the group-level observations of each phase separately so that cross-sectional observations in each test are independent of one another, although a given group's observations are not independent across phases. Markussen *et al.*

inclined subjects self-selected into use of the IS regime in the 6-C Treatment (see Appendix Table B.13). Our findings are thus consistent with the possibility that allowing groups to choose their institutions is efficiency-enhancing in our experiment in part because it lets heterogeneous individuals self-select into institutions that suit them better.[26] However, this need not rule out the possibility of a direct effect of voting, for instance that group members took others' votes for IS as an indication of intention to punish free riding.

i) How does IS perform when counter-punishment is possible?

Summary: Facilitating counter-punishment by letting subjects know who punished them does not undermine the effectiveness of IS in our experiment.

An important issue regarding the efficacy of informal sanctioning schemes was raised by Nikiforakis (2008), who pointed out that subjects in most experimental IS treatments are shielded from being punished back by those whom they punish due to absence of feedback about who punished whom by how much. When punished subjects are fully informed about the punishment they receive from each other group member and are given a distinct and immediate opportunity to counter-punish, Nikiforakis finds that earnings losses from punishing rise and that first-order punishment of free riders falls, thus eliminating the pattern of rising contributions usually associated with IS. This raises the possibility that the surprising popularity of IS despite availability of FS in our experiment would be sharply diminished if subjects could counter-punish in the IS regime.

To address this issue, we conducted two additional treatments that parallel the 6-C and endogenous IS comparison treatments except that (a) group members are provided with permanent identities and (b) information about their most recent and also earlier past punishments of one another is available alongside individual contribution information at each punishment stage (see the two "fuller information" treatments listed in the outer-right column of Table 1, and for treatment details, the Appendix). We find that the efficiency-

(forthcoming) also find contributions to be significantly higher in endogenously chosen than in exogenously imposed IS. In their data, the difference is statistically significant at the 5% level in a two-tailed test.

[26] There are differences as well as similarities to the results in Zhang *et al*. (2013). There, subjects choose between IS and a more centralized type of "pool punishment" by voting with their feet, similar to subjects in Gürerk *et al*. (2006), whereas in our majority voting setup a minority must live with the preferred institution of a group's majority. Zhang *et al*.'s pool punishment resembles efficiently targeted FS in that it targets all free riders and calls for payment up front even if no free riding occurs, but differs from FS in that contributions to the punishment pool remain individual decisions.

enhancing properties of the informal sanctions scheme are if anything strengthened by the inclusion of the additional feedback. Also, comparing the fuller information treatments having exogenous versus endogenous choice of sanctioning scheme when the alternative is costly FS, we find again that the IS scheme is popular and that contributions and earnings are significantly higher under IS when chosen by vote than when assigned exogenously, with the difference again being driven at least in part by subject self-selection. While our results indicate that allowing counter-punishment does not by itself reduce the efficiency-promoting potential of IS, it is important to note that there are differences in the way we introduce opportunities to counter-punish and the design of Nikiforakis (2008), so that our findings must be interpreted with care.[27] Details are given in Appendix C.

5. Discussion and Conclusion

Unlike the prediction from standard theory, the large majority of subjects in our experiment voted to use the formal sanctions mechanism only when it carried a low (here, zero) fixed cost. They selected instead the informal sanctions regime when formal sanctions were more costly. Votes for the informal sanctions regime were *ex post* rational in that subjects succeeded in reaching high levels of efficiency when informal sanctions were available. Informal sanctioning occurred sufficiently often and in a sufficiently well-targeted manner that in their presence, it became payoff-maximizing to contribute at least the observed group median level to the public good, and average contribution rose as low contributors risked being targeted with punishment. Our data confirms many qualitative results of Markussen *et al*. but for environments in which subjects themselves determine

[27] Differences include that (a) Nikiforakis's subjects have a stage in which counter-punishing is the only available activity immediately follows first-order punishing, whereas our subjects can only counter-punish in the following or later periods using the same punishment stage as is used for reacting to contributions; (b) Nikiforakis's subjects learn only of punishments aimed at themselves and thus their only higher-order punishment option is counter-punishment, whereas our subjects learn of all bi-lateral punishments in their group, so what Denant-Boemont *et al*. (2007) call "punishment enforcement" is also possible; and (c) Nikiforakis's subjects' identities are scrambled each period, whereas our subjects' identities remain visible, so higher-order punishment can take place over the course of several periods. Difference (a) may reduce the degree of counter-punishment in a "hot state," as one referee put it, or may reduce "experimenter demand" for counter-punishment, as Kamei and Putterman (2012) suggest. With respect to difference (b), our design resembles the "full information" treatment in Denant-Boemont *et al*. The effects of differing ways of making opportunities for higher-order punishment available to subjects are explored by both of the latter papers. Kamei and Putterman find that a treatment closely resembling that of Nikiforakis (2008) is the only one of six higher-order punishment treatments that achieves lower efficiency than treatments with first-order punishment only. However, they do not explore treatments that permit long sequences of dedicated counter-punishment stages, as do studies of "feuding" (e.g., Nikiforakis and Engelmann, 2011).

by voting the terms of the formal sanctions used, some subjects experience the available institutions before voting, formal sanctions carry both variable and fixed costs, and the formal sanction options are more neutral because perverse formal sanctions are always an option. The congruence of results is also remarkable since our subject pool is drawn from a diverse student body in a different country.

We found evidence that subjects' varying orientations towards cooperation, and in some cases their differing intelligence levels, mattered for their votes. More intelligent subjects were significantly more likely to vote for whichever of FS or IS was associated with higher earnings *ex post*. More conditionally cooperative subjects voted for more efficient FS parameters, and subjects inclined to punish perversely under IS were more likely to vote for less efficient FS parameters, including penalizing contributions to the public rather than private account. The fact that groups' majorities (and median voters) usually selected relatively efficient institutions and FS terms despite subject heterogeneity supports a "behavioral public choice theorem" first articulated in Ertan *et al*. (2009), namely that an unsung virtue of majority rule is that in social dilemmas, strongly anti-social or uncooperative inclinations are rarely shared by groups' majorities, so voting helps to neutralize them.[28]

Our demonstration that individuals achieve high levels of cooperation at low cost using informal sanctions, even when formal sanctions of modest cost are available, can be argued to be one of the strongest pieces of evidence for the presence and impact of the willingness to incur a private cost to punish in a large and growing experimental literature on the topic. Our results occurred despite the fact that we advantaged FS by modeling them as being imposed with certainty when triggered by the actions specified, whereas most real-world sanction schemes are fallible, both with respect to catching and imposing penalties on violators, and also the possibility of mistakenly punishing non-violators.

On the other hand, our design also advantaged IS by assuming that subjects always had cost-free and accurate information on which to condition their punishing. More realistically, accurate targeting of informal sanctions might depend on costly monitoring

[28] To be sure, uncooperative majorities may be found in some settings, as the results of Herrmann *et al*. (2008) suggest. How voting outcomes differ across societies is an interesting topic for future research.

choices by individuals, and lowered efficiency of IS due to imperfect observations (see Ambrus and Greiner, 2012) might tip preferences back towards formal sanctions. Future experiments could study subjects' willingness to pay to improve monitoring and reduce errors, as well as the effect of uncertainty of penalization and errors on choice of scheme. Another potentially complicating issue is that it may be more difficult in practice to prevent individuals from engaging in informal sanctioning activities than our design assumes, so the relevant choice might be formal sanctions accompanied by whatever informal sanctions individuals also decide to engage in (see Kube and Traxler, 2011 as well as related robustness treatments in Markussen *et al*.). Finally, the relative effectiveness of informal and formal sanctions in groups of larger scale might differ from what we observe in five-person groups, helping to explain the preference for formal sanctions at societal levels.

Ultimately, we think it useful to view our results from a perspective that sees voluntary or informal cooperation not simply as an alternative but also as a complement to formal authority and sanctions. Whereas availability of a formal sanctioning capacity that is responsive to majority will is automatic in our experiment, creating and sustaining such a capacity may in the real world require voluntary participation, monitoring, and other forms of engagement by citizens. Experimental evidence of voluntary collective action, including willingness to punish non-cooperators, might thus be seen not as indicating that the state is unnecessary, but as reassuring us that the cooperative underpinnings of the state are not out of reach. We hope to explore the interconnections between informal and authority-backed cooperation in future research.

References

Ambrus, Attila and Ben Greiner, 2012, "Imperfect public monitoring with costly punishment - An experimental study," *American Economic Review*102 (7): 3317-3332.

Andreoni, James, 1988, "Why Free Ride? Strategies and Learning in Public Goods Experiments," *Journal of Public Economics* 37(3): 291-304.

Bochet, Oliver, Talbot Page and Louis Putterman, 2006, "Communication and punishment in voluntary contribution experiments," *Journal of Economic Behavior and Organization* 60(1): 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.

Charness, Gary and Matthew Rabin, 2002, "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics* 117: 817-869.

Chaudhuri, Ananish, 2011, "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature," *Experimental Economics* 14(1): 47-83.

Cinyabuguma, Matthias, Talbot Page and Louis Putterman, 2006, "Can Second-Order Punishment Deter Perverse Punishment?" *Experimental Economics* 9(3): 265-279.

Dal Bó, Pedro, Andrew Foster and Louis Putterman, 2010, "Institutions and Behavior: Experimental Evidence on the Effects of Democracy," *American Economic Review* 100(5): 2205-2229.

Denant-Boemont, Laurent, David Masclet and Charles Noussair, 2007, "Punishment, Counter-punishment and Sanction Enforcement in a Social Dilemma Experiment," *Economic Theory* 33(1): 145-167.

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(5): 495-511.

Falk, Armin, Ernst Fehr and Urs Fischbacher, 2005, "Driving Forces Behind Informal Sanctions," *Econometrica* 73(6): 2017-2030.

Fehr, Ernst and Simon Gächter, 2000, "Cooperation and Punishment," *American Economic Review* 90(4): 980-994.

Fehr, Ernst and Klaus Schmidt, 1999, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics* 104: 817-868.

Fischbacher, Urs, 2007, "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics* 10(2): 171-178.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr, 2001, "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment," *Economics Letters* 71(3): 397-404.

Fischbacher, Urs and Simon Gächter, 2010, "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments," *American Economic Review* 100(1): 541-556.

Gächter, Simon and Benedikt Herrmann, 2008, "Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment," *Philosophical Transactions of the Royal Society B* doi:10.1098/rstb.2008.0275.

Gächter, Simon, Elke Renner and Martin Sefton, 2008, "The Long-Run Benefits of Punishment," *Science* 322(5907): 1510.

Gürerk, Őzgür, Bernd Irlenbusch and Bettina Rockenbach, 2006, "The Competitive Advantage of Sanctioning Institutions," *Science* 312(5770): 108-110.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter, 2008, "Antisocial Punishment across Societies," *Science* 319(5868): 1362-1367.

Kamei, Kenju, 2013, "Democracy and Resilient Pro-Social Behavioral Change: An Experimental Study," available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1756225.

Kamei, Kenju and Louis Putterman, 2013, "In Broad Daylight: Fuller Information and Higher-Order Punishment Opportunities Can Promote Cooperation," Brown University Department of Economics Working Paper 2012-3, Revised.

Kube, Sebastian and Christian Traxler, 2011, "The Interaction of Legal and Social Norm Enforcement," *Journal of Public Economic Theory* 13(5): 639-660.

Markussen, Thomas, Louis Putterman and Jean-Robert Tyran, forthcoming, "Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes," *Review of Economic Studies* (in press).

Nikiforakis, Nikos, 2008, "Punishment and Counter-punishment in Public Goods Games: Can we Really Govern Ourselves?" *Journal of Public Economics* 92(1-2): 91-112.

Nikiforakis, N., Engelmann, D., 2011. Altruistic Punishment and the Threat of Feuds. *Journal of Economic Behavior and Organization* 78, 319–332.

Nikiforakis, Nikos and Hans-Theo Normann, 2008, "A Comparative Statics Analysis of Punishment in Public Goods Experiments," *Experimental Economics* 11(4): 358-369.

Nosenzo, Daniele and Martin Sefton, 2012, "Promoting Cooperation: the Distribution of Reward and Punishment Power," Center for Decision Research & Experimental Economics Discussion Paper 2012-08, University of Nottingham.

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(4): 495-521.

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.

Page, Talbot, Louis Putterman and Bulent Unel, 2005, "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency," *Economic Journal* 115(506): 1032-1053.

Putterman, Louis, Jean-Robert Tyran and Kenju Kamei, 2011, "Public Goods and Voting on Formal Sanction Schemes: An Experiment," *Journal of Public Economics* 95(9-10): 1213-1222.

Sigmund, Karl, Hannelore De Silva, Arne Traulsen and Christoph Hauert, 2010, "Social Learning Promotes Institutions for Governing the Commons," *Nature* 466: 861 – 863.

Sutter, Matthias, Stefan Haigner and Martin Kocher, 2010, "Choosing the Carrot or the Stick? – Endogenous Institutional Choice in Social Dilemma Situations," *Review of Economic Studies* 77(4): 1540-1566.

Thöni, Christian, Jean-Robert Tyran and Erik Wengström, 2012, "Microfoundations of Social Capital," *Journal of Public Economics* 96(7-8): 635-643.

Traulsen, Arne, Röhl Torsten and Manfred Milinski, 2012, "An Economic Experiment Reveals that Humans Prefer Pool Punishment to Maintain the Commons," *Proceedings of the Royal Society* B 279: 3716-3721.

Tyran, Jean-Robert and Lars P. Feld, 2006, "Achieving Compliance when Legal Sanctions are Non-deterrent," *Scandinavian Journal of Economics* 108(1): 135-156.

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

Zhang, Boyu, Cong Li, Hannelore De Silva, Peter Bednarik and Karl Sigmund, forthcoming, "The Evolution of Sanctioning Institutions: An Experimental Approach to the Social Contract," *Experimental Economics* (in press).

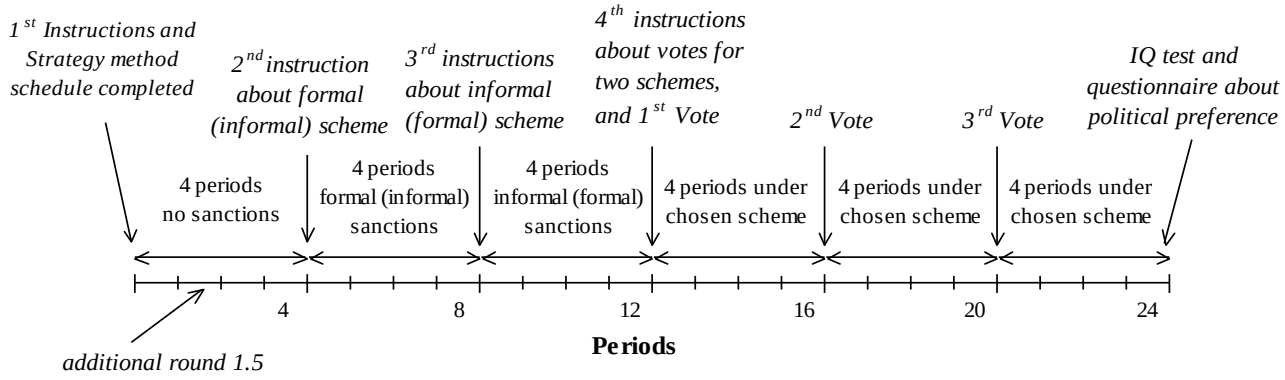
***Table 1*:** Summary of treatments, sessions, and group and subject numbers

		No Fixed Administrative Cost of FS (N)	Fixed Administrative Cost of FS (C)	
3-Vote Treatments	NS→IS→FS	**Treatment: 3(IF)-N** Sessions: 2 Groups: 8 Subjects: 40	**Treatment: 3(IF)-C** Sessions: 2 Groups: 8 Subjects: 40	
	NS→FS→IS	**Treatment: 3(FI)-N** Sessions: 2 Groups: 7 Subjects: 35	**Treatment: 3(FI)-C** Sessions: 2 Groups: 8 Subjects: 40	
6-Vote Treatments		**Treatment: 6-N** Sessions: 2 Groups: 7 Subjects: 35	**Treatment: 6-C** Sessions: 2 Groups: 8 Subjects: 40 **Exogenous IS Comparison Treatment** Sessions: 2 Groups: 6 Subjects: 30	**6-C "Fuller Info" Variant** 2 7 35 **Exog IS "Fuller Info" Variant** 2 8 40
		Treatment: Baseline (Exogenous-NS) Sessions: 2 Groups: 8 Subjects: 40		

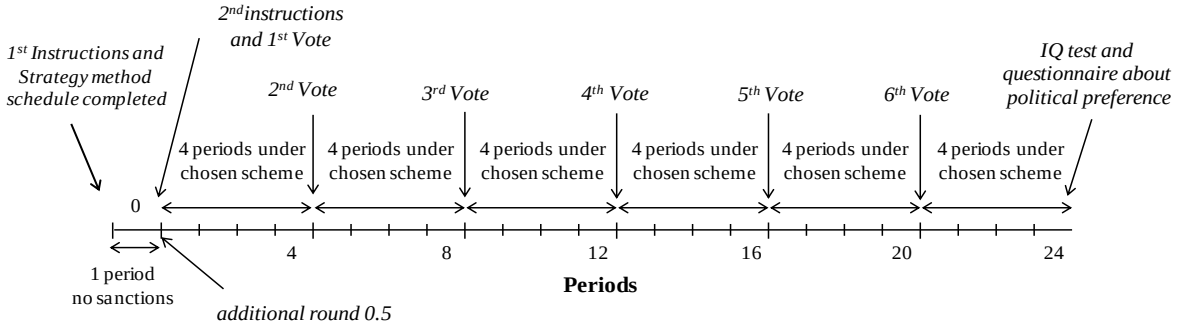
Notes: NS = no sanctions, FS = formal sanctions, IS = informal sanctions. The experiment as a whole consisted of 20 sessions in which 75 groups composed of 375 individual subjects participated. There were a total of 188 group votes on the use of FS versus IS, with 958 individual votes on that question. The Exogenous IS comparison treatment (dashed rectangle) appears in the cell of Treatment 6-C because its parameters and structure are identical to that of groups selecting IS in 6-C treatment except for the fact that the IS scheme is imposed rather than chosen by vote. For time structures of 3-Vote, 6-Vote and Baseline treatments, see Figure 1.

***Figure 1*:** Experimental Design

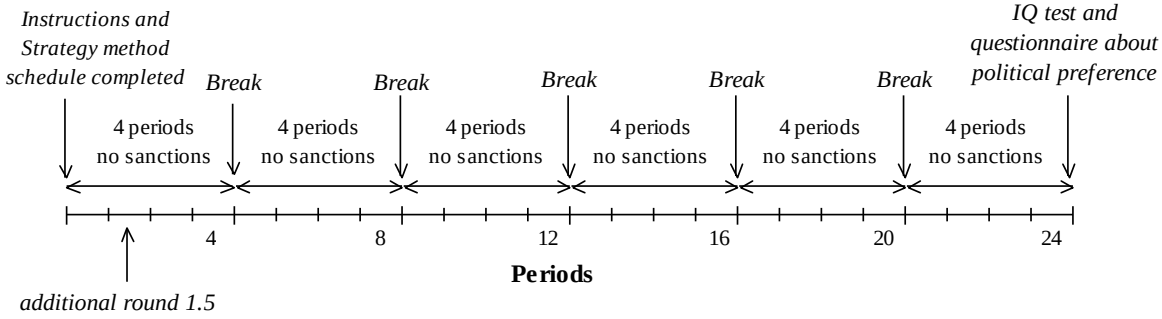
(A) 3-Vote treatments



(B) 6-Vote treatments

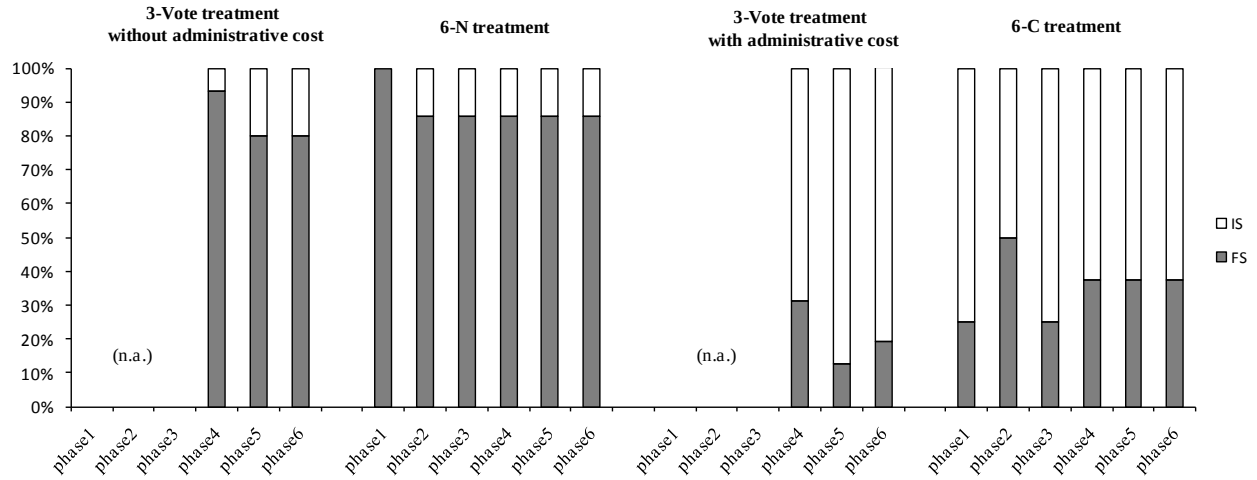


(C) BASELINE treatment

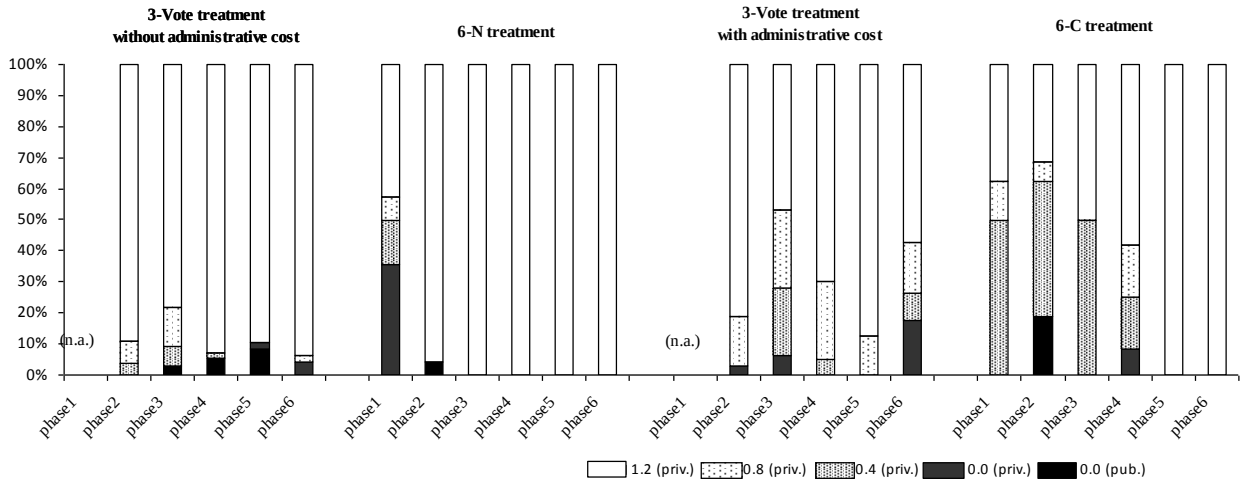


Note: The additional rounds 1.5 in 3-Vote and Baseline treatments and 0.5 in 6-Vote treatment, are the periods in which the conditional contribution schedules described in footnotes are drawn upon for one randomly selected group member to determine all group members' payoffs.

***Figure 2*:** Share of voting outcomes for formal vs. informal sanctions, and for sanction rates, by phase



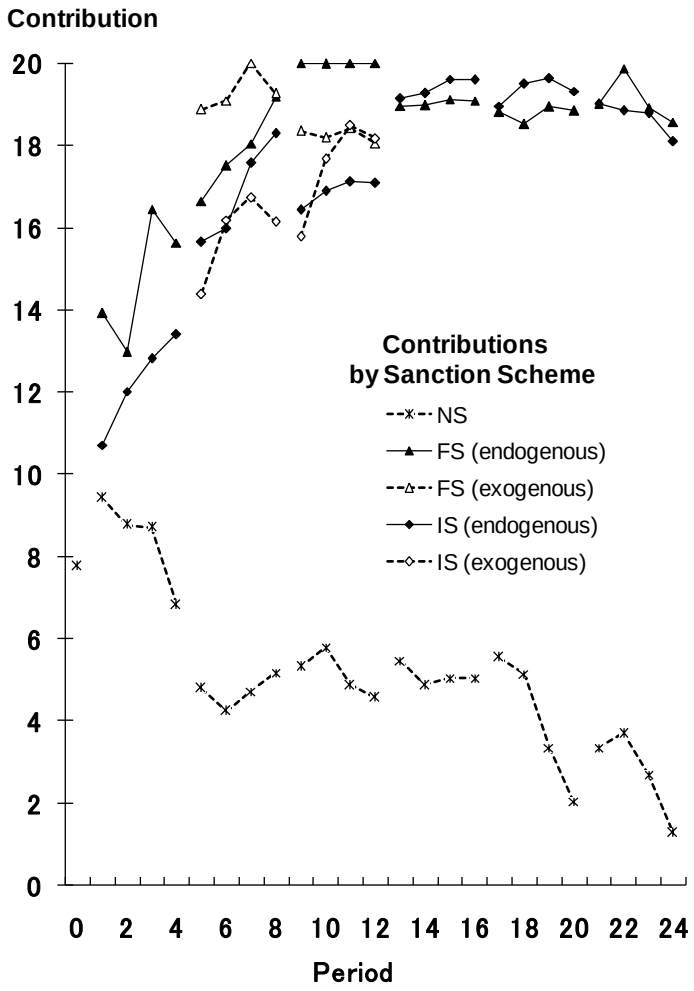
(a) Share of formal vs. informal scheme vote outcomes, by phase



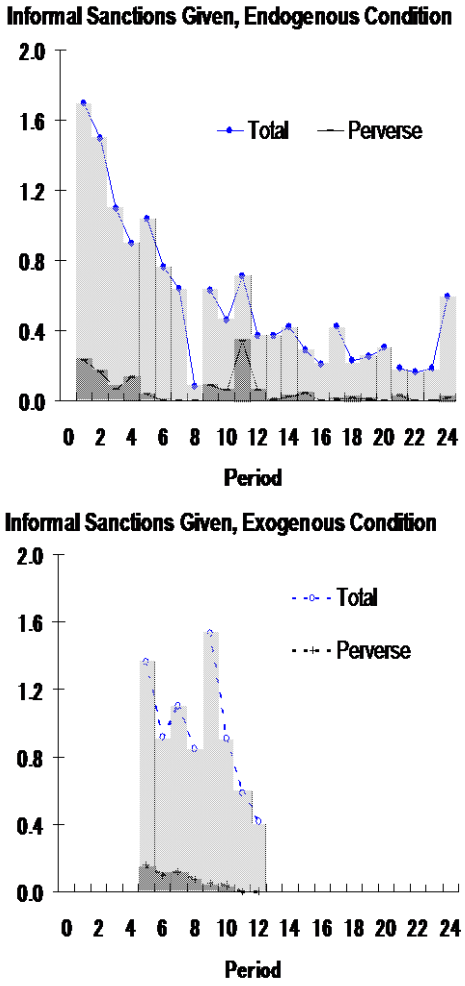
(b) Shares of voting outcomes for penalty rates, by phase

Note: data are group outcomes, not individual votes.

***Figure 3*:** The trends of average contribution to the public account and average amount of informal sanctions given

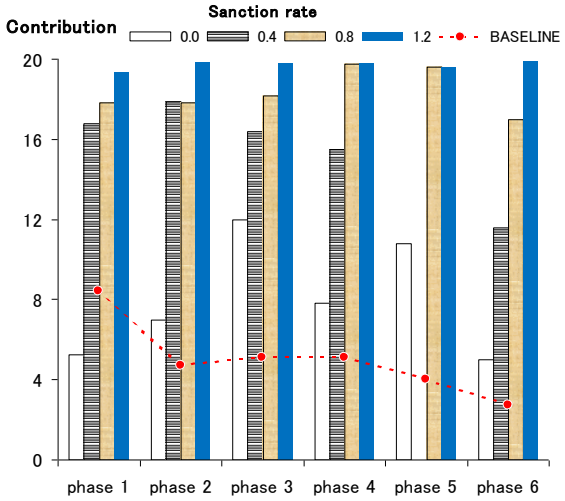


(a) Average Contribution



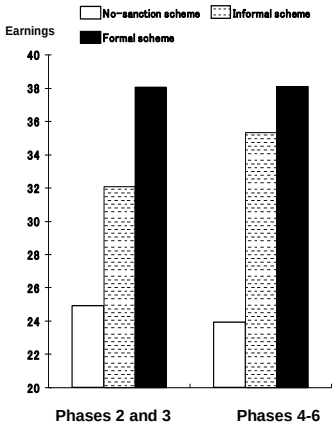
(b) Average Amounts of Informal Sanctions Given

***Figure 4*:** Contributions under formal scheme by sanction rate

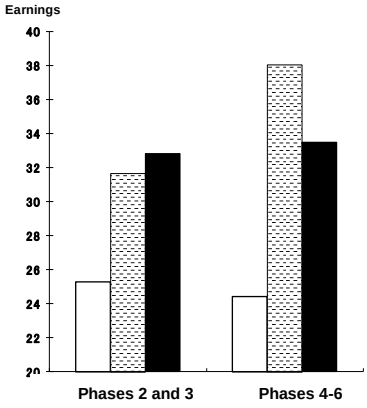


Note: No group chose a sanction rate of 0.4 in phase 5.

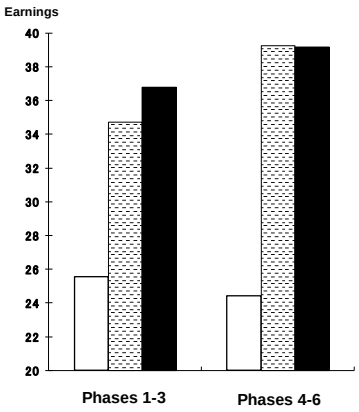
***Figure 5*:** Average earnings under NS, IS, and FS in early and late phases



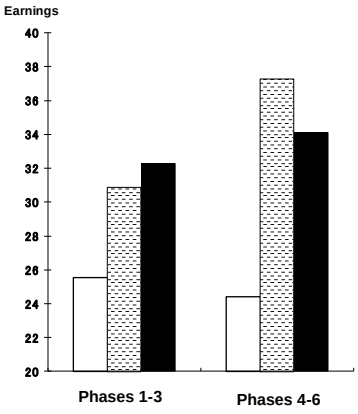
(a) 3-Vote without administrative cost



(b) 3-Vote with administrative cost



(c) 6-N treatment



(d) 6-C treatment