Markussen, Thomas; Putterman, Louis G.; Wang, Liangjun

Working Paper

Governing collective action in the face of observational error

Working Paper, No. 2017-2

Provided in Cooperation with:
Department of Economics, Brown University

Suggested Citation: Markussen, Thomas; Putterman, Louis G.; Wang, Liangjun (2017) : Governing collective action in the face of observational error, Working Paper, No. 2017-2, Brown University, Department of Economics, Providence, RI

This Version is available at:
https://hdl.handle.net/10419/202614

Governing Collective Action in the Face of Observational Error

Thomas Markussen, Louis Putterman and Liangjun Wang*

August 2017

Abstract: We present results from a repeated public goods experiment where subjects choose by vote one of two sanctioning schemes: peer-to-peer (informal) or centralized (formal). We introduce, in some treatments, a moderate amount of noise (a 10 percent probability that a contribution is reported incorrectly) affecting either one or both sanctioning environments. We find that the institution with more accurate information is always by far the most popular, but noisy information undermines the popularity of peer-to-peer sanctions more strongly than that of centralized sanctions. This may contribute to explaining the greater reliance on centralized sanctioning institutions in complex environments.

JEL codes: H41, C92, D02

Keywords: Public goods, sanctions, information, institution, voting

* University of Copenhagen, Brown University, and Zhejiang University of Technology, respectively. We would like to thank The National Natural Science Foundation of China (Grant No.71473225) and The Danish Council for Independent Research (DFF|FSE) for funding. We thank Xinyi Zhang for assistance in checking translation of instructions, and Vicky Zhang and Kristin Petersmann for help in checking of proofs. We are grateful to seminar participants at University of Copenhagen, at the Seventh Biennial Social Dilemmas Conference held at University of Massachusetts, Amherst in June 2017, and to Attila Ambrus, Laura Gee and James Walker for helpful comments.

Introduction

When and why do groups establish centralized, formal authorities to solve collective action problems? This fundamental question was central to classical social contract theory (Hobbes, Locke) and has recently been the subject of a group of experimental studies (Okada, Kosfeld and Riedl 2009, Traulsen et al. 2012, Andreoni and Gee 2012, Fehr and Williams 2013, Zhang et al. 2014, Markussen, Putterman and Tyran 2014, Kamei, Putterman and Tyran 2015, Gross et al. 2016, Nicklisch, Grechenig and Thöni 2016). The present paper contributes to this literature by investigating how collective choice between formal and informal institutions is affected by the information environment, in particular by the frequencies with which inaccurate signals raise the danger of misdirected punishment. We implement a public goods experiment in a laboratory and let groups vote about whether to adopt peer-to-peer sanctions á la Fehr and Gächter (2000, 2002) or a centralized authority, which automatically imposes sanctions on free riders, but comes with a fixed cost (cf. Markussen, Putterman and Tyran 2014). We vary the quality of information about contributions to the public good, which is available to peers and to the central authority, respectively. In one treatment, the central authority has access to more accurate information than peer punishers, in another treatment the reverse is true, and in the remaining treatments, the quality of information is the same in both institutions (high quality in one treatment, low quality in another).

Markussen, Putterman and Tyran (2014) study collective choice between formal (centralized) and informal (peer-to-peer) sanction systems and find considerable support for informal sanctions against a formal sanctions alternative, even when the fixed cost of formal sanctions is low and the sanctions are strong enough to deter free riding according to standard theory. However, information is never noisy in that study or in the related experiment of Kamei, Putterman and Tyran (2015), in which sanction scheme parameters are also determined by voting. Ambrus and Greiner (2012) consider the effect of noisy information in an exogenously imposed environment with peer-to-peer sanctioning and find that a moderate amount of noise (a 10 percent probability that a public goods contributor is reported to peers as a free rider) leads to sharply lower earnings, relative to the situation with perfect information. Similar findings are reported by Grechenig, Nicklisch and Thöni 2010. Markussen, Putterman and Tyran (2016) study the effect of punishment errors in formal sanctioning and find that such errors undermine the deterrent effect of sanctions significantly. None of the studies investigate choice between formal and informal institutions in the presence of noise, but their results motivate the main hypothesis of the present study, namely that the information environment is likely to have a significant effect on choice of institutions.

Two important results emerge from the paper, one largely intuitive, the other somewhat less so. The first result is that, consistent with the main hypothesis, even though we only introduce moderate amounts of

noise (a 10 percent risk that a contribution to the public good is reported incorrectly, which does not alter the equilibrium predictions under standard economic theory), we find strong treatment effects on the choice of institutions. Whereas in our baseline treatment with perfect information in both formal and informal institutions, about half of groups choose formal sanctions (much as in Markussen *et al*.), in the treatment in which information is noisy in the informal but not in the formal sanction environment, the share choosing formal sanctions rises to more than 90 percent. Conversely, in a treatment in which information is noisy with formal but not with informal sanctions, only about 30 percent of groups choose formal sanctions. When both formal and informal sanctioning is affected by noise, about 70 percent choose formal sanctions, substantially more than in the baseline treatment.

The just-mentioned shifts towards whichever scheme operates on better information, when only one environment is noisy, are consistent with intuition. But the fact that that shift is asymmetric—more strongly toward (an almost 40 percentage point increase) than away from (only about a 20 percentage point reduction of) formal sanctions when there is noise in one environment—is not as intuitive. Combining this with the bias (a roughly 20 percentage point jump) towards formal sanctions, when both environments are noisy, we get our second major result: imperfect information undermines the popularity of informal sanctions more strongly than it undermines the support for formal sanctions. A potential implication is that increasing complexity of the social environment, e.g. due to increasing community size or openness, might act as a force favoring the adoption of formal institutions, even if information problems afflict peer and top-down observation equally.

The strong treatment effects on voting for institutions are partly explained by the fact that there are also significant treatment effects on contributions to the public good, and hence on earnings. Earnings are close to their maximum feasible level under formal sanctions without error, and are not significantly lower under informal sanctions without error. But whereas introducing observational errors lowers earnings in the formal sanctions environment by an average of about 12%, errors lower average earnings under informal sanctions to a larger degree, about 20%. The greater earnings difference is a result of punishment being not only less accurately targeted when information is noisy, but also of the apparent reluctance to punish by some (perhaps inequality averse) subjects, in an informal sanctions regime. On the other hand, treatment effects on voting remain significant even when earnings in earlier phases of the experiment under each type of institution are controlled for in regressions. This suggests that people are averse to imperfect information (and/or to mis-targeted punishment), for reasons which go beyond its effect on material pay-offs. A possible interpretation, we suggest, is that if rules must be enforced but with unavoidable occurrences of error, then many may prefer an impersonal vertical structure over peer-to-peer

enforcement, since the latter but not the former seems burdened by unwitting "betrayal" and "moral scruples" issues. We expand on this explanation of why bias away from peer enforcement may emerge in the face of error in our predictions discussion.

Arguably, our findings are important because decisions about whether to introduce centralized, formal sanctioning institutions are taken at a number of levels. Firms choose between horizontally organized production teams (i.e. peer monitoring) and hierarchical structures of management (supervisors on shop floors, superiors and subordinates in lines of organizational authority). Users of common property resources, such as communal forests or water sources, decide on whether to control resource use by informal peer monitoring, or by establishing a central authority. In such cases, the state also decides whether to intervene or to leave resource management to the users themselves. Resource users may encourage or resist state intervention. Countries decide on whether to solve international, collective action problems, such as pollution or refugee crises, through bilateral arrangements or by surrendering sanctioning power to a supra-national body, such as the European Union or the World Trade Organization. The results of this experiment suggest that such institutional choices may be sensitive to the quality of information available to actors with more local versus global vantage points.

One example is the regulation of fisheries. In her seminal book on common property governance, Elinor Ostrom (1990) describes three different inshore fisheries in Turkey (Alanya, Bodrum and Izmir). One of these, Alanya, successfully maintained fishing activity at sustainable levels through self-governance. The two others did not. One notable feature of the regulatory system devised by the fishermen in Alanya is that it allowed them to monitor each other's activities closely and at low cost (each was allocated a certain area of the bay to fish; fishing outside one's allotted area was likely to be detected by those whose area one was encroaching upon). Hence, the quality of information was high. Such systems were not in place in the other two fisheries, in part because they are larger. Ostrom mentions several potential reasons for the failures of self-regulation in Bodrum and Izmir (e.g. heterogeneous outside options), but it seems plausible that information problems played an important role. In general, the regulation of fisheries also exemplifies how the quality of information available to peers and to central authorities may vary. In a small-scale fishery such as Alanya (with approximately 100 local fishers) peers may well have high-quality information about each other. Conversely, it may be excessively costly for a government agency to attempt to monitor such small communities effectively, if geographically separated and marked by local idiosyncrasies. In a larger scale fishery, such as Izmir (1,800 fishers), exploiting a larger expanse of the sea, fishermen may find it much more difficult to monitor other fishermen. On the other hand, to regulate production at such scale, it might be feasible for the government to set up systems of, for example, monitoring the movement of boats

and meticulously registering the catches of fish landed in ports. Conceivably the associated costs may even be covered by the fishermen, e.g. through licensing fees. Hence, depending on the circumstances, either peers or central regulators may have an information advantage. The results of the present study support the idea that relative informational advantages significantly affect the choice of whether to regulate fishing through peer monitoring or centralized regulation.

Our framework is also potentially relevant for understanding why the state has gradually assumed responsibility for regulating various types of anti-social behavior once regulated mainly by local communities. One example is parents' use of corporal punishment against children, which in many countries has become increasingly subject to legal sanctions. The common interpretation seems to be that this results from a general reduction in society's tolerance of violence. An alternative or supplementary interpretation is that in traditional, small-scale societies, neighbors had extensive information about each other and would sanction excessive use of violence against children. In the modern city, on the other hand, such informal sanctioning is rare, and the state therefore needs to step in.[1] Similar explanations can be offered for increased formalization of the regulation of (for example) littering, snow clearing and building standards.

However, insofar as we find that noise tends to bias choice towards centralized enforcement even when equally present at the center and among peers, our results suggest that something beyond the balance of informational advantages may also be at play. We suggest behavioral explanations of this bias which rest on the idea that impersonality is preferred over personal interactions conflicted by "moral scruples" and potential sense of "betrayal," if enforcement errors are unavoidable. To return to one of our examples, it may be best that someone prod each resident to clear the snow from their walk, but neighbors who lack in depth knowledge of one another may prefer not to complain when there might be valid excuses. They may opt, instead, to leave neighbor-to-neighbor relations as they are, with (sometimes unfair) enforcement actions given over to city hall. Assigning authority on some matters to governing bodies might likewise help to reduce potential bi-lateral tensions between states, and such arrangements may be more likely to be opted for the more are the relevant facts potentially subject to informational problems.

The paper is organized as follows. Section 2 discusses related papers and section 3 presents the experimental design. Section 4 discusses theoretical predictions and section 5 presents the results. Section 6 concludes.

[1] Jared Diamond (2012, pp. 194-195) gives the example of the linguist Daniel Everett, who lived among the Piraha Indians in the Amazonian rainforest and was prevented from spanking his daughter by his neighbors, who disapproved of corporal punishment! In the jungle village, spanking one's daughter without being observed by the neighbors was difficult, indeed.

1. Related papers

A number of papers have investigated the role of imperfect monitoring and/or uncertainty in social dilemmas. For example, Palfrey and Rosenthal (1994) investigated a repeated public goods game in which marginal rates of substitution between the public and the private good varied across individuals and were private information in some treatments. Results showed that imperfect information reduced contributions to the public good. Cason and Khan (1999) implemented a public goods game in which subjects in some treatments could only observe each other's behaviors once every six rounds of play (as opposed to receiving feedback at the end of each round). This type of limited information did not affect levels of cooperation. Carpenter (2007) presents results from a public goods game with peer-punishment, in which subjects can only monitor and punish a certain fraction of their fellow group members. He found that levels of cooperation increased with the degree of observability. Fudenberg, Rand and Dreber (2012) implemented a repeated prisoners' dilemma game, where the action implemented ("cooperate" or "defect") differs from the intended action (i.e. the action chosen by the subject) with a certain probability. This design differs from ours in the sense that errors occur in the *execution* of strategies, rather than their *observation*. They found that strategies are more "lenient" (i.e. the tendency to defect in response to defection by the other player is lower) when the error rate is positive than when it is zero. Aoyagi and Frechette (2009) investigated imperfect monitoring (rather than imperfect execution) in an infinitely repeated prisoners' dilemma game, finding that noisy information reduced pay-offs. Aoyagi, Bhaskar and Frechette (2015) analyzed a similar setting but distinguished between "perfect" monitoring, and two types of noisy monitoring, labeled "public" and "private". With "public monitoring," subject *i* sees a noisy signal about subject *j*'s action and also learns which signal *j* received about *i*'s action. With "private monitoring," *i* receives a signal about the behavior of *j* but does not observe the signal *j* sees about *i*. In this terminology, the present paper compares "perfect monitoring" with noisy "public information" (i.e. subjects do learn whether their actions were reported correctly to others or not). Aoyagi, Bhaskar and Frechette find that cooperation levels are comparable in all three conditions, but that strategies are more lenient with noisy, private monitoring than with perfect monitoring.

As mentioned in the Introduction, Ambrus and Greiner (2012) is an important inspiration for the present paper because it finds a strong effect of imperfect monitoring on efficiency in a public goods game with peer punishment. Ambrus and Greiner (2015) revises the design of Ambrus and Greiner (2012) by adding what they call a "democratic punishment" condition. In this condition, subjects in each period state for each fellow group member whether they should be punished or not (only one level of positive punishment

is possible). Subjects are punished if a majority of other group members state that they should be. They find that efficiency is higher with democratic than with "individual" (peer-to-peer) punishment, but also that noise reduces efficiency significantly with both individual and democratic punishment. Fischer, Grechenig and Meier (2016) study a similar set-up, but instead of the "democratic punishment" condition in Ambrus and Greiner (2015), they implement a treatment where the power to punish is concentrated in the hands of a randomly selected individual. They find that such centralization does not increase efficiency and that noisy information is detrimental to subject earnings with centralized as well as with decentralized punishment, although "perverse punishment" (punishment directed at cooperators) is less frequent with centralized than with decentralized punishment. Our study differs from Ambrus and Greiner (2015) and from Fischer, Grechenig and Meier (2016) both in the nature of the centralized punishment regime and, most significantly, in that we study endogenous choice of institutions.

Arguably, the paper most closely related to ours is Nicklisch, Grechenig and Thöni (NGT, 2016), which also studies choice between centralized and decentralized institutions and varies the quality of information between treatments. There are, however, a number of important differences between their study and ours. First, the central authority in NGT is a subject in the experiment who chooses which others to punish and how much, whereas in ours this role is played by an automaton (the computer). This difference mirrors a difference of approach in the literature, with some papers, such as Yamagishi (1986), Okada, Kosfeld and Riedl (2009), Putterman, Tyran and Kamei (2011), Andreoni and Gee (2012), Markussen, Putterman and Tyran (2014) and Kamei, Putterman and Tyran (2015), joined by the present paper, having the authority be fully automated (written into the experimental software) and others, including Gürerk et al. (2009), Heijden et al. (2009), Traulsen et al. (2012), Fehr and Williams (2013), Zhang et al. (2014), Nosenzo and Sefton (2014), Gross et al. (2016), Fischer, Grechenig and Meier (2016), and NGT, having the role of central authority be played by an experimental subject. Our central authority always punishes according to the signal it receives, regardless of whether information is noisy, whereas our peer punishers are free to react differently to noisy than to perfect information, their decisions thus being of interest to investigate. The discretionary nature of centralized punishment in NGT and its automatic nature in our experiment places our papers on opposite ends of the design spectrum, with our own approach representing the "impersonal" feature of authority to which we refer in later discussion. Probably, most real-life authorities are somewhere in between these extremes, enjoying some amount of discretion but also being subject to some formal or informal constraints generated by constitutions, social norms, or concern about re-election or potential dismissal or recall. Many times authorities also are guided or constrained by explicit rules, sometimes ones selected by the same agents the authority is charged with disciplining.

Second, NGT never vary the *relative* quality of information available to centralized and decentralized punishers, respectively, while this study does so. We believe this is a realistic and important contribution, since information gathering by a network of peers and by a centralized authority, respectively, is often based on quite different technologies, as explained in the fisheries example above. Peers may rely on personal observations and bilateral communication, while a centralized authority would tend to use formalized and perhaps technologically advanced reporting and monitoring procedures. Depending on circumstances, one type of technology may deliver more accurate information than the other. It is important to investigate whether such differences in the accuracy of information affect the relative popularity of institutions.

Third, subjects in NGT "vote with their feet", i.e. they choose individually which institution to join, whereas our subjects vote by "casting ballots," with a simple majority in a group of fixed size determining which institution is used. The voting approach corresponds to a mechanism by which institutional choices are often made and has different properties, for instance that a minority may well be overruled by the majority.[2] Other significant differences between NGT and our study include quite different error structures, and availability in NGT of a sanction free environment as an option additional to centralized and decentralized punishment institutions.

2. Experimental design

Public goods game and sanctioning institutions

We consider the following, basic environment. Groups of size *n* play a public goods game with binary contribution decisions.[3] The game is repeated a number of times and groups are fixed ("partner matching"). In each period, each individual receives an endowment of size *W* and decides whether to keep this endowment or allocate it to public goods production. Individual *i's* pay-offs, π_i, are calculated according to equation 1:

$$\pi_i = W - C_i + a\sum_{j=1}^{n} C_j \qquad (1)$$

[2] Ertan et al. (2009), Putterman et al. (2011) and Hauser et al. (2014) discover, in experimental settings, cases in which pro-social majorities overrule anti-social minorities when choosing institutions, rules, or policies in social dilemmas.
[3] To test the robustness of our results, we also report below a variant using a range of intermediate contribution options, a more common set-up in the literature. We explain there the advantage for current purposes of the binary approach, but report most qualitative results to be similar.

where $C_j \in \{0, W\} \forall j$ and $1/n < a < 1$. The latter condition ensures that contributing to the public good is collectively but not individually optimal, i.e. that the group faces a social dilemma.

We introduce two different types of sanctioning institutions in this environment, namely informal sanctions (IS) and formal sanctions (FS).

With *informal sanctions*, subjects observe whether each other group member contributed to the public good or not. They can then decide to reduce the earnings of other group members, at a cost to themselves (Ostrom, Gardner and Walker 1992, Fehr and Gächter 2000, 2002). The information on contribution decisions is anonymized, such that individuals cannot be tracked from period to period. As in the above and most related experiments, punished individuals are also not informed which member or members punished them.[4] Pay-offs are calculated as follows:

$$\pi_i^{IS} = W - C_i + a\sum_{j=1}^{n} C_j - \sum_{j \neq i} p_{ij} - \sigma \sum_{j \neq i} p_{ji} \tag{2}$$

where p_{ij} is the number of *punishment points* used by subject *i* to reduce the earnings of subject *j* and p_{ji} is the number of punishment points *i* receives from *j*. σ is the punishment effectiveness factor, i.e. the reduction in earnings for each punishment point received.

With formal sanctions, a central authority automatically sanctions free riders by deducting *sW* points from their earnings. To maintain the central authority, each individual must pay a fixed fee equal to *c* each period.[5] Earnings are calculated as follows:

$$\pi_i^{FS} = \left(W - C_i\right)\left(1 - s\right) + a\sum_{j=1}^{n} C_j - c \tag{3}$$

Timeline and voting

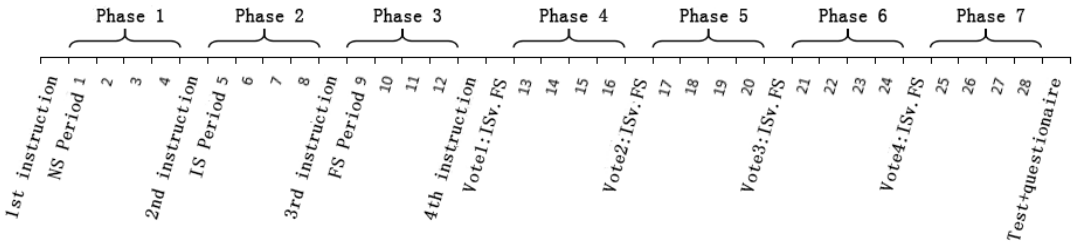
The experiment is divided into seven phases with four periods in each. The first phase is simply four periods of the standard public goods game with binary contributions, i.e. earnings are given by equation 1. In Phase 2, all groups play with IS and in Phase 3 all groups play with FS. At the beginning of Phase 4, groups vote

[4] Although investigating robustness to this aspect may be an important issue for future research, it seems best to begin with features common to most past research. We revisit the issue in our Conclusions.

[5] Markussen *et al*. (2014) suggest that incurring a social cost of punishment up front (by paying for police, courts, prisons, etc.) versus incurring a private cost only when punishing occurs may be a core difference between the real world counterparts of FS and IS. Kamei *et al*. (2015) adopt a fixed cost of FS for the same reason; although they include also a variable cost of FS when actual sanctions occur, their qualitative findings are nonetheless quite similar.

about whether to use IS or FS. Voting is compulsory and the majority decides. The outcome of the vote is implemented in all four periods of the phase. This procedure is repeated in phases 5, 6 and 7.

Figure 1 Timeline



Our main focus of analyses is the voting outcomes. The reason for introducing three phases with exogenous institutions before voting begins is the desire to ensure that voters have a thorough understanding of the nature of the game and the alternatives they can vote for. Arguably, voting results are most interesting when such an understanding is established. The reason for introducing IS in Phase 2 and FS in Phase 3 (rather than the other way around, or in a random sequence) is that this is arguably a natural sequence. Since punishment emerges endogenously in IS, it is likely that punishment patterns in IS would be altered if subjects were exposed to FS before IS, because the experience of FS would frame perceptions about who should be punished, and who not. On the other hand, since who gets punished in FS is exogenous, it will not be affected by prior exposure of subjects to IS. Having multiple periods within phases provides us with evidence of adjustments and learning under given institutions.[6]

Imperfect information and treatment variation

A key feature of the experiment is that in some treatments, peers or the central authority or both do not receive perfect information about contribution decisions. In particular, when there is imperfect (noisy) information, there is an *x* probability that any given contribution is reported wrongly to peers in IS, and/or to the central authority (as well as to peers) in FS. So, if a group member contributed *W* to the public good, there is an *x* percent chance that this is reported as 0. We call this "type I error." If he or she contributed 0,

[6] A possible design issue is whether FS should *replace* IS or instead be introduced *alongside* IS (Kube and Traxler 2011, Markussen, Putterman and Tyran 2014). The argument in favor of the latter design is that even when centralized sanctioning institutions are in place, peer-to-peer sanctioning such as ostracism and bad-mouthing still take place. On the other hand, it is also true that a key role for central authorities such as states is to eliminate the more serious forms of vigilante activities (lynching, vendettas etc.). We feel that both views have merit but we choose to implement a design in which FS replaces rather than complements IS since it has the advantage of greater simplicity and because Markussen et al. do not find large qualitative differences in results, in their robustness treatments.

there is an *x* percent chance that this is reported as *W*. We call this "type II error."[7] This set-up is similar to that used by Ambrus and Greiner (2012), except that subjects in their experiment only face a risk of type I errors, not type II errors. We include type II errors because in many cases this type of error is equally or more common than type I errors and because the rules are still reasonably simple to explain to subjects.

When there is noisy information, earnings in IS are still given by equation 2. This means public goods production is still based on true contributions, not on the possibly erroneous reports of contributions. Of course, earnings may be indirectly affected by imperfect information if noise affects punishment decisions, which in turn affects contributions. The results in Ambrus and Greiner (2012) and Grechenig, Nicklisch and Thöni (2010) indicate that this effect is potentially strong.

In FS, earnings under imperfect information are given by:

$$\pi_i^{FS,noise} = W - C_i + a\sum_{j=1}^{n} C_j - s\left(W - C_i^{reported}\right) - c \qquad (3')$$

where $C_i^{reported}$ is the contribution value reported to peers and the central authority. So, the central authority always punishes according to the signal it receives, regardless of whether that signal is potentially affected by noise.

We implement four treatments, defined by the combination of two binary variables, namely a) whether information in IS is noisy or not and b) whether information in FS is noisy or not. Table 1 presents the resulting four treatments. These include two treatments in which quality of information is the same for both institutions (NoNoise, NoiseBoth), and two treatments in which quality of information varies between institutions (NoiseIS and NoiseFS).

Table 1 Treatments

		Noisy information in IS	
		No	*Yes*
Noisy information in FS	*No*	NoNoise	NoiseIS
	Yes	NoiseFS	NoiseBoth

[7] As Markussen et al. (2016) note, the literature varies regarding what to designate as type I versus type II error. Our usage here parallels the designations in that paper—the error that may lead to "punishing the innocent" is type I, the one that may "let the guilty go free" is type II. But we here associate errors with observations, rather than with punishments, because although "erroneous punishment" automatically follows from erroneous observation in the case of FS, that is not necessarily the case with IS.

Parameter values and feedback

Parameter values are set taking results in the existing literature into account and with the aim of creating realistic expectations that a) both IS and FS improve efficiency relative to the standard voluntary contributions mechanism (VCM) when information is perfect, and b) voting between the two institutions is approximately balanced in the NoNoise treatment, an outcome that gives scope for detecting the impact of errors on the relative popularities of the two institutions. Several parameter values match those in Markussen, Putterman and Tyran (2014). We always set the groups size n to 5, the endowment W is always 20 and the marginal per capita return to the public good, a, is always 0.4. These have been common parameter values in public goods experiments. We set the punishment effectiveness σ to 4 because Nikiforakis and Normann (2008) report that within four periods of play (the duration of a phase in our experiment), informal sanctions increase earnings relative to a sanction free environment only if punishment effectiveness is at least 4, and we are interested in studying sanctioning institutions that potentially increase efficiency. s, the level of sanctions in FS, is set to 0.8, implying that subjects reported as free riders pay a fine of 0.8*20 = 16 points. This means that sanctions are deterrent, both with and without noise in FS (see below). The fixed cost of FS, c, is 3. Markussen, Putterman and Tyran (2014) found, with other parameter values identical to those chosen here, that a bit less than 60 percent of subjects chose FS when c = 2, whereas a large majority chose IS with c = 8. Accordingly, we anticipate that setting c = 3 might generate approximately even voting between IS and FS in the treatment without noisy information. The probability of erroneous information, x, is either 0 or 0.1, depending on treatment and institution (x = 0.1 corresponds to Noise in treatment Table 1). So, when information is noisy, it means that there is a 10 percent chance that any given contribution is reported incorrectly. This is also the level of noise in the noisy treatments of Ambrus and Greiner 2012. Other papers, e.g. NGT, Markussen, Putterman and Tyran (2016) and Fischer, Grechenig and Meier (2016) have used x values up to 0.5. In that light we believe it is fair to describe the level of noise in the present experiment as "moderate."

In terms of feedback, participants learn at the end of each period what the true, total amount of contributions to the public good was. They also learn whether they were themselves exposed to a type I or a type II information error, and they learn how many others in their group were affected by type I and type II errors, respectively. Subjects are informed about the total amount of punishment they have received, and about their final earnings for the period. After each vote, subjects learn the outcome of the vote but not the precise distribution of votes.

Treatments with continuous contribution decisions

In addition to the four treatments described above, where contribution decisions are binary, we implemented a parallel set of treatments which were essentially identical, except that contribution decisions were quasi-continuous. In particular, contributions to the public good could be any integer value between 0 and *W* (both values included). In these treatments, imperfect information has the following structure: For any contribution higher than 0, there is an *x* percent risk that the contribution is reported to peers and to the central authority as 0 (type I error). For any contribution lower than *W*, there is an *x* percent risk that the contribution is reported to the others as *W*. This error structure with quasi-continuous contributions was also used in Markussen, Putterman and Tyran (2016). The reason to run both treatments with binary and with continuous contribution decisions is that both approaches offer distinct advantages. The error structure is somewhat easier to explain in the binary treatments, and the set-up is comparable to that in Ambrus and Greiner 2012, a closely related paper. On the other hand, most public goods experiments with or without punishment (Ledyard, 1995, Zelmer, 2003, Fehr and Gächter, 2000, Chaudhuri, 2010), including many of the most closely related papers (Markussen et al., 2016, Grechenig et al. 2010, NGT, forthcoming, Fehr and Williams, 2013), have used continuous contribution decisions, and the continuous contribution choices arguably offer the possibility of observing richer dynamics in the repeated public goods game than in the game with binary contributions. The main drawback of a (quasi-)continuous design is that if type I errors entail falsely reporting positive contributions as 0 and if some subjects are expected to refrain from punishing reported 0 contributors in IS because such errors are known to occur, low contributors (who might otherwise have contributed, say, 3 of 20 points) have an incentive to contribute 0, a value behind which they can to some extent "hide." The potential interplay between the temptation thus created for low contributors and the possible amendment of punisher behavior creates challenges to the analyst that can be avoided by adopting the binary approach.[8] Despite this, the extent of distortions created by the "hiding behind zeroes" problem is modest and the results from the two sets of treatments are similar. We focus mostly on results from the treatments with binary contributions, but results from the continuous treatments are presented in the appendix and commented on in the main text.

[8] Other alternatives, not pursued here, would be to retain the continuous set-up but model errors themselves as potentially multivalued (say normally or uniformly distributed around the true contribution) or as random values (as in Grechenig et al. 2010 and Nicklisch et al. forthcoming.

Implementation

The experiment was conducted at the Economics and Management Laboratory at Zhejiang University of Technology in Hangzhou, China, during the fall of 2015 and the spring of 2016.[9] A total of 16 experimental sessions were conducted, eight for the treatments with binary contribution decisions and eight for the treatments with continuous contributions. There were 30 participants in each session, except one session of the NoNoise treatment with binary contributions, which had 25 participants. A total of 475 subjects participated. Subjects earned on average 64 Yuan (about 9USD), including a 10 Yuan show-up fee.[10] Subjects received written instructions at the beginning of phases 1, 2, 3 and 4, read along as the experimenter read the instructions aloud, and answered on-screen comprehension questions in each of those instances. Instructions and on-screen messages were first written in English, then translated to Chinese, and then translated back to English by independent experts. The experiment was implemented with the software z-Tree (Fischbacher 2007). Subjects were recruited by posting notices around campus.

3. Predictions

Predictions based on standard, economic theory are straightforward. First, since punishment in IS is costly, nobody engages in punishment in the last period of play, even when there is perfect information. Therefore, no one contributes to the public good either. By backward induction, this logic applies to all periods. The introduction of noisy information does nothing to change this equilibrium. Hence, predicted

[9] The choice of location is based simply on this being the home institution of one of the researchers. The university is located in one of China's most developed coastal cities, Hangzhou, in Zhejiang province, and experiment participants were drawn from many disciplines. 48 percent of participants were from liberal arts majors, the rest from various other programs. 53 percent of participants were women. There appears to be little reason for concern that subjects in China behave differently than those in the Western countries in which most existing experiments of this type have been based. Supplementary material in Fu, Ji, Kamei and Putterman (2017), who conducted a voluntary contributions with punishment experiments at Nankai University in Tianjin, China, at approximately the same time, shows that both contribution and punishment behaviors there, as well as at the Chengdu, China site in the cross-country experiment of Herrmann, Thöni and Gächter (2008), are well within the range of behaviors observed at the latter's Western sites, i.e. Bonn, Boston, Copenhagen, Melbourne, Nottingham, St. Gallen and Zurich. Comparison of the decisions in Phase 2 of our Error Free, continuous contribution treatment in Hangzhou shows contributions and punishments to be qualitatively quite similar to those in the Chengdu and Tianjin samples, as well. We thank Tingting Fu for assisting with this comparison. Finally, while none of the comparisons just mentioned is from an experiment that includes voting on institutions, the results we present below for the NoNoise treatment suggest that setting parameters with the aim of inducing voting behaviors based on subject behaviors of Markussen *et al*. (conducted in Copenhagen) is quite successful in achieving our intended goal of a roughly equal institutional split, suggesting quite similar behavioral propensities among Hangzhou and Copenhagen students.

[10] The exchange rate of 1 point = ¥0.055 was set based on estimated point earnings, estimated session duration of about 90 minutes, and the estimate that students can earn about ¥35 - ¥45 per hour in other part-time employment. The achieved average payout was consistent with these considerations.

earnings per subject in all periods, with or without noisy information, is *W*. In FS, sanctions are "deterrent" (i.e. they turn contributing to the public good into the privately optimal strategy) in the noise-free environment if $s > (1 - a)$. Since we use $s = 0.8$ and $a = 0.4$, this condition is fulfilled. When noisy information is introduced, it can be shown that sanctions continue to be deterrent as long as $x < (s - (1 - a))/2s = 0.125$ (See Appendix A). Since we use $x = 0.1$, sanctions are theoretically deterrent even when information is noisy. Expected earnings are $anW - xsW - c = 40 - 1.6 - 3 = 35.4$, which is higher than *W* (the predicted earnings under IS) for the parameters used here. So, the introduction of noisy information does not change predicted contributions in either IS or FS and predicted earnings are always highest in FS. Hence, income maximizing voters should always vote for FS in all treatments.

Things are considerably more complicated when behavioral theory, and results from previous experimental studies, are taken into account. First, the standard theory prediction on the effects of IS tends to be highly inaccurate. As shown by Fehr and Gächter (2000, 2002) and a number of other papers, including Markussen *et al*. (2014), IS in many cases leads to very high levels of contributions to the public good. On the other hand, Ambrus and Greiner (2012) and Grechenig, Nicklisch and Thöni (2010) show, as discussed in the introduction, that noisy information severely undermines the efficiency of IS.

One prominent model of social preferences is Fehr and Schmidt's theory of "inequity aversion" (Fehr and Schmidt 1999). The model assumes the following utility function:

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

where α_i and β_i are parameters measuring agent *i*'s aversion to disadvantageous and advantageous inequality, respectively (other notation as above). Fehr and Schmidt show that if agents have this utility function, a cooperative equilibrium may exist in a public goods game and that such an equilibrium is more feasible when peer punishment is possible than when it is not (Fehr and Schmidt 1999, propositions 4 and 5). Here we present a version of Fehr and Schmidt's Proposition 5, which is augmented to take noisy information into account. The aim is to investigate how noisy information affects the feasibility of a cooperative equilibrium, and the properties of such an equilibrium. The setting is the IS version of the public goods game, described above. Denote by $b_i(x)$ the probability assigned by enforcer *i* to a fellow group member being a cooperator, given that he or she is reported as a free rider. The appendix shows that $b_i(x)$ is increasing in *x*, as long as *i*'s prior beliefs assign a strictly positive, prior probabilities to any fellow group member being a cooperator and to any fellow group member being a free rider. It also shows that

$b_i(x) = 0$ when $x = 0$. Assume that players know the distribution of social preferences in their group, but that information about contributions to the public good is presented in such a way that they cannot link observed contribution decisions to the social preferences of the individuals making the decisions.[11]

Proposition 1 *(augmented Fehr and Schmidt)*: Suppose there is a group of n' "conditionally cooperative enforcers", $1 \le n' \le n$, with preferences that obey $a + \beta_i > 1$ and

$$\sigma > \frac{(n-1)(\alpha_i + 1) - (n'-1)(\alpha_i + \beta_i) - \alpha_i}{\alpha_i - b_i(x)(\alpha_i + \beta_i)} + 1 \qquad (5)$$

for all $i \in \{1, \ldots, n'\}$

whereas $\alpha_i = \beta_i = 0$ for all $i \in \{n'+1, \ldots, n\}$. Also assume that

$$0 \le x \le \frac{1}{2} \quad \text{and} \quad x \le \frac{1}{\sigma + 3 - n} \qquad (6)$$

Then the following strategies, which describe the players' behavior on and off the equilibrium path, form a subgame perfect equilibrium.

- In the first stage each player contributes $C_i = W$
- If no player is displayed to the others as contributing $C_i = 0$, there are no punishments in the second stage. If any player, *i*, is displayed to the others as choosing $C_i = 0$, then each enforcer $j \in \{1, \ldots, n'\}$, $j \neq i$, chooses p_{ji} such that $\sigma p_{ji} = \dfrac{W}{n' - \dfrac{1}{\sigma} - \left(2n' - \left(1 + \dfrac{3-n}{\sigma}\right)\right)x}$ while all other players do not punish. (Proof in Appendix A).

Based on the proposition, we can distinguish between two different channels through which observational errors affect earnings under IS:

[11] This assumption ensures that players never conclude with certainty that a contribution decision is displayed correctly or incorrectly (when $x > 0$). It is roughly analogous to our experimental design, where subjects have information about the social preferences of fellow group members from earlier periods of the experiment, but the presentation of information about contributions to the public good is anonymized.

First, the proposition shows that even when noise is present, a cooperative equilibrium may exist. However, earnings in such an equilibrium are decreasing in *x* (the probability that a contribution choice is displayed incorrectly to fellow group members) because each time a contributor is wrongly displayed to the others as a free rider, costly punishment is handed out. That is, the incidence of "perverse punishment" (punishment directed toward contributors) increases with the frequency of observational error, which in turn reduces earnings and leads to inequality.[12]

Second, if we assume that enforcers assign strictly positive, prior probabilities for all other players, *j*, to both $C_j = W$ and $C_j = 0$ (i.e. we assume that enforcers are a priori not 100 percent certain about which strategies other group members will choose), then equation (5) shows that conditions for the existence of a cooperative equilibrium are tougher to meet when the probability of observational error is higher (i.e. the right-hand side of the equation is increasing in *x*, because $b_i(x)$ is increasing in *x*, given the assumptions made, see appendix). The reason is the following: Punishment of free riders is driven by aversion to disadvantageous inequality (α_i). However, when observational errors may occur, there is a chance that a punisher inadvertently punishes a cooperator, which leads to advantageous inequality in favor of the punisher. Hence, if the punisher is sufficiently averse to advantageous inequality, he or she may hesitate to punish, for fear of committing a type I error. Note that conditionally cooperative enforcers have $\beta_i > 1 - a$ by definition, so all enforcers are at least somewhat averse to advantageous inequality.[13] Noting the Oxford English Dictionary's definition of the term as "a feeling of doubt or hesitation with regard to the morality or propriety of a course of action," we call the reluctance to punish out of fear of committing type I error the "scruples" factor.[14]

---

[12] The concepts of "perverse" and of "antisocial" punishment are overlapping, but we use the former because our subjects cannot know whether their punisher is a cooperator, so only the contribution status of the punishment recipient is relevant. See Fu and Putterman (2017) for disambiguation of the two concepts and some comparisons in application to experimental data analysis.

[13] Formally, the punisher gains utility equal to $\frac{1}{n-1}\alpha_i(\sigma - 1)$ from each point of punishment directed toward a free rider, but loses $\frac{1}{n-1}\beta_i(\sigma - 1)$ from punishing a contributor. Hence, the expected gain is $\frac{(\sigma-1)}{n-1}\left((1-x)\alpha_i - b_i(x)\beta_i\right)$, ignoring the monetary cost of punishment and changes in inequality toward other group members. This is decreasing with *x*, and more so when β_i is high than when it is low.

[14] Assumption (6) also imposes restrictions on the amount of noise compatible with cooperative equilibria. The institution behind the assumption is that if *n* and *x* are large, relative to *σ*, then the cost paid by enforcers for

Note that, for the kind of equilibrium set out in Proposition 1 to hold, the amount of punishment received by free riders must be increasing in *x* for all feasible parameter values. The intuition is that since punishment is not perfectly targeted because of type I as well as type II errors, the punishment must be correspondingly harsher in order to be deterrent. This may contribute to explaining the finding of "punishment despite reasonable doubt" in Grechenig, Nicklisch and Thöni (2010). In contrast, subject' "scruples" nudge outcomes towards less, not more, punishing in the presence of errors.

Factors additional to inequity aversion are also likely to play a role. Bohnet and Zeckhauser (2004), Bohnet et al. (2008), Aimone and Houser (2011) and Aimone et al. (2014) find that many people display so-called "betrayal aversion" when deciding whether to trust another. Such individuals prefer a risk of failure of given probability when governed by chance alone to the same chance of betrayal by a trusted human interaction partner. Aversion to the possibility of being punished by a fellow group member despite having cooperated may exist for similar reasons. To be sure, in the trust or investment game on which the cited authors focus (see also Berg et al. 1995), a second mover's failure to reciprocate is more straightforwardly a case of betrayal of trust than is a cooperator's being punished when displayed as a defector, in our setting. Nonetheless, being punished by an interaction partner despite knowing that the report of defection could be an error may generate similar feelings of betrayal,[15] and the contrast between such punishment under IS and punishment in FS as an impersonal result of a random process may resemble the contrast discussed by them. We hypothesize that this distinction might influence choice between informal and formal sanctions in a setting of imperfect information.

Consider the NoiseBoth treatment, in which there is a risk of wrongful punishment under formal as well as under informal sanctions. Whereas in IS the agent of wrongful punishment is an individual person and fellow group member, in FS that agent is an impersonal authority. If subjects are more averse to being wrongfully punished by other individuals than by an impersonal mechanism, then this leads to the prediction that noisy information undermines the popularity of IS more than the popularity of FS. Outside of the lab, of course, punishment by central authorities is implemented by persons, not by computers, but

---

punishing cooperators wrongly displayed as free riders becomes too large to make enforcement profitable. The assumption is easily met in our set-up, where $x = 0.1$ and $1/(\sigma+3-n) = ½$.

[15] As a particularly simple example, suppose (abstracting from Type II errors) that IS with 10% Type I error probability is underway and that group members are being reported over several periods as contributing about 90% of the time. Then members might infer that reported non-contribution events are with high likelihood instances of error. An individual punished by a fellow group member when reported, in such a situation, might feel betrayed. Clearly, more judgment is entailed here than in the trust game case, so extension of the "betrayal aversion" concept to our environment requires considerable qualification. "Aversion to failing to receive the benefit of the doubt" is a more accurate term for our case, so the reader should understand that our use of the existing "betrayal aversion" terminology is a convenience, with the term applying in a more general sense than in that of the original literature.

conceivably many people experience the actions of formal authority as emerging from an impersonal "system," rather than from an individual, and the psychological reaction to sanctions imposed by central, bureaucratic authorities is therefore likely to be different from the reaction to sanctions imposed by neighbors, colleagues or other peers who have the option to desist from punishing you given their uncertainty about your guilt.

The "scruples" issue mentioned above can also affect voting on institutions beyond the effect through earnings, paralleling the operation of "betrayal aversion" (in our extended sense of that term). Although subjects with strong scruples may avoid engaging in punishment when IS is in place and information is noisy, many may be willing to overcome their scruples and engage in punishing if necessary, yet would suffer psychological discomfort from doing so. Having to participate in IS with noise, and thus to "wrestle with their scruples," will be a direct source of disutility to such individuals, and the possibility of avoiding this by instating FS can work either alongside of or instead of (extended) "betrayal aversion" as a factor tipping the scale towards voting for FS. Of course, favoring FS for this reason requires that the logical role of one's own vote in putting in place the "impersonal" system that sometimes perversely punishes would need to be overlooked or assigned little weight. However, viewing FS as largely impersonal despite having voted for it may be fairly common, given that voting on institutions is a rare event whereas wrestling with whether to punish or not would be an almost constant ongoing issue under IS with noise.

In sum, when behavioral factors are taken into account, cooperative equilibria in IS may exist, which opens up the possibility that subjects may prefer IS over FS, as observed in many groups in Markussen *et al*. (2014). Preference for IS over FS should if anything strengthen when information is noisy in FS and not in IS.[16] However, our behavioral analysis also points to three reasons for expecting IS to be less popular when information is noisy in IS than when it is not: (1) greater incidence of punishment directed at cooperators (perverse punishment); (2) hesitation to punish seeming free riders due to the danger of perversely punishing cooperators (scruples); and (3) desire to avoid being subjected to "betrayal" by others and/or to avoid suffering the psychological cost to oneself of "wrestling with scruples." The first of these factors affects FS as well as IS, whereas the last two factors apply only to IS. While the first two factors affect

[16] Note that in case there is a cooperative equilibrium in IS, earnings (according to otherwise standard theory) are higher under IS than under FS, because there is a fixed fee in FS and not in IS. However, whereas in the theory actual punishment is unnecessary, in reported experiments high cooperation in the presence of punishment opportunities require that there first be considerable real punishment. In relatively successful cases, that is, punishment declines with repetition after the threat of it has been rendered credible by its exercise, but we know of no cases in which the hypothetical threat alone suffices. Presence of such punishment is perhaps explicable by reference to uncertainty about others' preferences.

institutional choice indirectly, through their effects on earnings under IS relative to FS, the last factor does not affect monetary earnings, but operates, rather, directly at the level of collective choice.

Hence, while standard economic theory predicts that observational errors do not affect institutional choice (FS should be extremely popular in all treatments), behavioral theory predicts IS is sometimes popular, and that both IS and FS are less popular when affected by error than when not. The behavioral theory laid out above also predicts that errors have a stronger effect on the popularity of IS than on the popularity of FS, since only factor (1) affects FS, while all factors affect IS, and that the effect of noise on choice of institutions will be partly but not fully mediated through earnings (since factors (1) and (2) operate through earnings but factor (3) does not).

4. Results

Since our main focus is the effect of imperfect information on choice of formal vs. informal institutions, we begin by presenting results on voting.

Voting

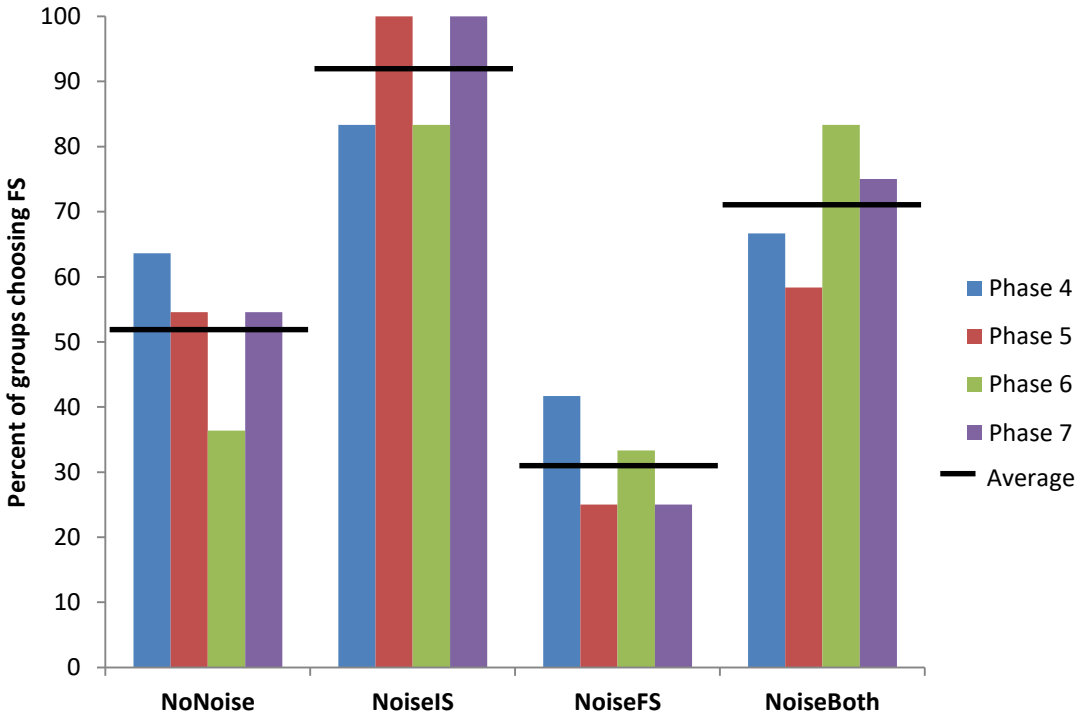
Figure 2 presents the main results of the paper, namely the average, group level voting outcomes by treatment and phase. The figure shows the share of groups choosing formal sanctions rather than informal sanctions. The colored bars show the mean in each phase, while horizontal, black lines show the overall averages. Table 2 presents Mann-Whitney tests of the pairwise treatment comparisons.

The results show strong treatment effects. We begin by noting that the share of groups choosing FS in the NoNoise treatment is close to 50 percent. As explained in section 3, parameters were selected using predictions based on results in Markussen, Putterman and Tyran (2014) with the aim of achieving such a starting point from which to assess the effects of introducing noise. The result is nevertheless remarkable, first, in suggesting that behavior among experimental subjects in China is reasonably predictable based on results from subjects in Denmark, where the experiment presented in Markussen, Putterman and Tyran was conducted. This strengthens the external validity of these results.[17] Second, the vote outcomes (along

[17] One cannot rule out that behavioral similarity is greater among students at competitive, cosmopolitan universities around the world than among, say, average working people in the respective societies. Relatedly, some anthropologists including Henrich et al. (2010) argue that university students' behaviors may poorly represent those among broader spectra of societies. There is considerable variation across world regions even for university student subjects (e.g., Herrmann et al. (2008)), however, so our observation is simply that in the kind of decision problem we study, there is no indication that university students in Hangzhou, China behave differently from western counterparts, and thus no reason to treat our results with greater caution than those of the more numerous

with the scheme performances reported below) reconfirm the competiveness of IS in an environment with perfect information, as found by both Markussen *et al*. and Kamei *et al*. (2015). Together, these results constitute an impressive body of evidence that people are sufficiently predisposed towards punishment of free riders so that, at least when information is good, informal sanctions can rival formal sanctions in effectiveness, contrary to the predictions of non-behavioral economic models. Recall that the formal sanction regime is theoretically expected to generate a surplus of 20 points, at a fixed cost of only three points, relative to the theoretical prediction of full free riding for IS.

Figure 2 Choice of institution by treatment and phase, group level



Note: N = 47 groups, observed four times each. 12 groups are observed in each treatment except NoNoise, which has 11 groups.

The results show strong treatment effects. We begin by noting that the share of groups choosing FS in the NoNoise treatment is close to 50 percent. As explained in section 3, parameters were selected using predictions based on results in Markussen, Putterman and Tyran (2014) with the aim of achieving such a starting point from which to assess the effects of introducing noise. The result is nevertheless remarkable,

experiments reported in international periodicals in which the subjects pools are students in Western Europe or its offshoots in North America and Oceania. See again the Appendix of Fu et al., 2017, for evidence that Chinese and Western university students' behaviors are more similar than those of some other university student subject pools studied by Herrmann et al.

first, in suggesting that behavior among experimental subjects in China is reasonably predictable based on results from subjects in Denmark, where the experiment presented in Markussen, Putterman and Tyran was conducted. This strengthens the external validity of these results.[18] Second, the vote outcomes (along with the scheme performances reported below) reconfirm the competiveness of IS in an environment with perfect information, as found by both Markussen *et al*. and Kamei *et al*. (2015). Together, these results constitute an impressive body of evidence that people are sufficiently predisposed towards punishment of free riders so that, at least when information is good, informal sanctions can rival formal sanctions in effectiveness, contrary to the predictions of non-behavioral economic models. Recall that the formal sanction regime is theoretically expected to generate a surplus of 20 points, at a fixed cost of only three points, relative to the theoretical prediction of full free riding for IS.

Table 2 Mann-Whitney tests of treatment effects on voting, group level

	Phase 4	Phase 5	Phase 6	Phase 7	Mean of voting in all phases	N
NoNoise vs. NoiseIS	0.29	0.01***	0.02**	0.01***	0.01**	23
NoNoise vs. NoiseFS	0.30	0.16	0.88	0.16	0.30	23
NoNoise vs. NoiseBoth	0.88	0.86	0.02**	0.33	0.33	23
NoiseIS vs. NoiseFS	0.04**	0.00***	0.02**	0.00***	0.00**	24
NoiseIS vs. NoiseBoth	0.36	0.01**	1.00	0.07*	0.04**	24
NoiseFS vs. NoiseBoth	0.23	0.11	0.02**	0.02**	0.02**	24

Note: M-W tests. Entries are p-values. Level of observation: groups.

Turning now to the treatments not previously studied at any site, results for the NoiseIS treatment show that the introduction of a moderate amount of noise in IS, while maintaining perfect information in FS, has a dramatic effect on voting. In this treatment, more than 90 percent of groups voted for FS. The difference between NoNoise and NoiseIS is highly significant in three out of four phases and overall (i.e. when the share of vote outcomes for FS, by group, is considered). Hence, there is strong evidence that having imperfect information among peers, only, undermines the popularity of peer-to-peer sanctions. Results for

[18] One cannot rule out that behavioral similarity is greater among students at competitive, cosmopolitan universities around the world than among, say, average working people in the respective societies. Relatedly, some anthropologists including Henrich et al. (2010) argue that university students' behaviors may poorly represent those among broader spectra of societies. There is considerable variation across world regions even for university student subjects (e.g., Herrmann et al. (2008)), however, so our observation is simply that in the kind of decision problem we study, there is no indication that university students in Hangzhou, China behave differently from western counterparts, and thus no reason to treat our results with greater caution than those of the more numerous experiments reported in international periodicals in which the subjects pools are students in Western Europe or its offshoots in North America and Oceania. See again the Appendix of Fu et al., 2017, for evidence that Chinese and Western university students' behaviors are more similar than those of some other university student subject pools studied by Herrmann et al.

the NoiseFS treatment show a weaker, although still economically significant effect of introducing noisy information in FS, relative to the NoNoise treatment. In this treatment, about 30 percent of groups chose FS, a drop of about 20 percentage points relative to NoNoise. This difference is not statistically significant according to Table 2's group level non-parametric tests. The fact that the effect of introducing noise in FS is only about half as large as the effect of adding noise to IS (20 vs. 40 percentage points, comparing with NoNoise) and that the difference between NoiseFS and NoNoise is not statistically significant, while the difference between NoiseIS and NoNoise is, supports the conclusion that noisy contribution information undermines the popularity of IS more strongly than it does the popularity of FS.

This conclusion is further strengthened by the comparison between NoNoise and NoiseBoth. Although the same degree of informational noise is added to both sanctioning institutions, in the latter treatment, the almost perfectly equal numbers of votes for FS and IS in NoNoise give way to a clear majority of vote outcomes (around 70%) favoring FS in NoiseBoth. While this difference is only statistically significant in one phase (Phase 6), it is economically substantial (20 percentage points), thus corroborating the view that imperfect information strengthens the popularity of FS relative to IS. In the two last phases of the experiment and overall, four out of six pairwise treatment comparisons are statistically significant, even though the number of observations is limited (23 or 24). This supports the general conclusion that the information environment has strong effects on institutional choice, even when only moderate amounts of noise are considered.

Table 3 presents regression analyses of voting outcomes. Regression (1) checks only for treatment effects, finding that FS is adopted significantly more often in NoiseIS than in the omitted NoNoise treatment, with statistically insignificant differences for NoiseFS and NoiseBoth. Since voting is preceded by phases with exogenous IS and FS (phases 2 and 3) and, later on, by phases with endogenous institutions (phases 4-6), it is possible that voting is driven by experiences with the alternative institutions in previous phases, particularly by the relative earnings generated in either institution. Potentially, treatment effects could be entirely generated by these experiences. To see if this is the case, regressions (2) and (3) add controls for earnings in previous phases with IS and FS, respectively. Regression 2 controls for earnings in the exogenous phases, Phase 2 and Phase 3, while regression 3 controls for earnings in the most recent phase a group used each institution prior to the vote, regardless of whether endogenously or exogenously.

Table 3 Voting regressions, group level

	Dependent variable: Group voted for FS		
	(1)	(2)	(3)
NoiseIS	0.394***	0.236**	0.188**
	(0.137)	(0.093)	(0.089)
NoiseFS	-0.21	-0.159	-0.114
	(0.171)	(0.104)	(0.083)
NoiseBoth	0.186	0.203*	0.167*
	(0.157)	(0.105)	(0.100)
Earnings in Phase 2 (IS)		-0.008***	
		(0.001)	
Earnings in Phase 3 (FS)		0.004***	
		(0.001)	
Earnings in most recent phase with IS			-0.009***
			(0.001)
Earnings in most recent phase with FS			0.005***
			(0.001)
Constant	0.523***	0.894***	0.85***
	(0.129)	(0.223)	(0.227)
R-sq (overall)	0.22	0.47	0.55
N	188	188	188

Note: Random effects, linear regressions. Level of observation: Groups. Standard errors clustered at group level. The reference treatment is NoNoise. *, ** and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Results show strong effects of earnings in previous phases, in the expected direction. However, the treatment effect of NoiseIS remains significant, although its size is reduced approximately by half. Moreover, the treatment effect of NoiseBoth goes from insignificant to significant at the 10 percent level when the earnings controls are introduced, with little change in the point estimate. Thus, both treatments that include errors under IS display treatment effects over and above those attributable to earnings alone—implying a dislike of IS with errors in and of itself. This supports the prediction that noisy information undermines the popularity of IS more strongly than the popularity of FS due to behavioral or psychological factors other than earnings—e.g., the desires to avoid "betrayal" and "wrestling with scruples" posited in Section 4. The "partial R-squared" of the treatment dummies in model 3, i.e. the share of variation in institutional choice which is explained by treatment effects above and beyond the effects working through income, is 11 percent. Comparing this with the standard R-squared value of 22 percent in model 1, which

includes only the treatment dummies, suggests that "betrayal aversion" and "wrestling with scruples" accounts for approximately half of the total effect of treatment differences.[19]

In summary, the results are in line with our theoretical predictions that a) errors would affect the popularity of both FS and IS, b) the effect would be stronger for IS than for FS and c) the effect of error would be partly but not fully mediated by earnings, because factor (3) in the theoretical framework (desire to avoid betrayal and the discomfort of scruples) does not operate through earnings. All these predictions are supported by the data.

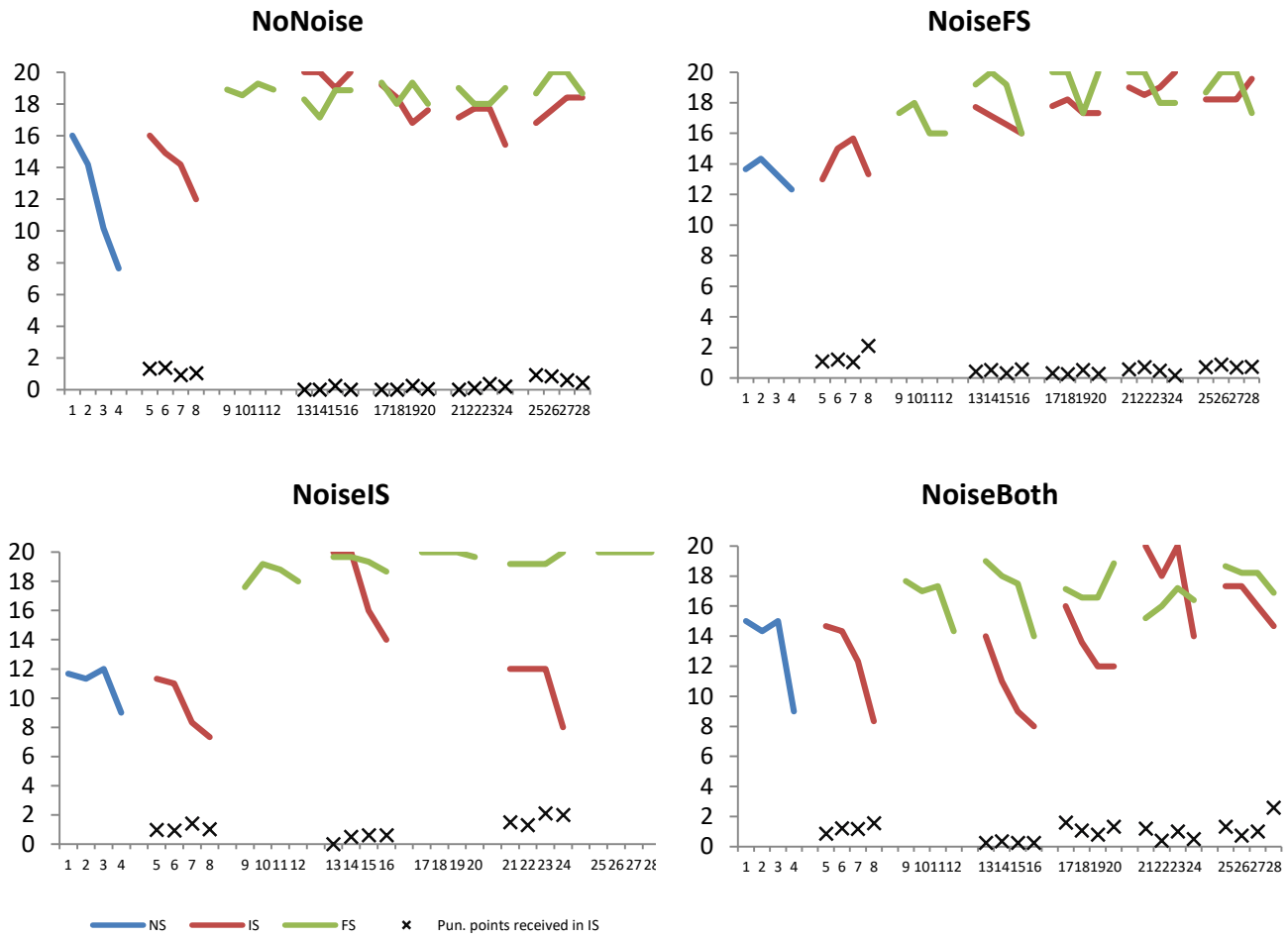
Results on voting in the treatments with continuous contribution decisions are shown in Figure C1 and Tables C2 and C3 in Appendix C. Findings are generally similar to those from the treatments with binary contribution decisions. The figure shows that here, too, the proportion of group votes for FS rises relative to NoNoise in both NoiseIS and NoiseBoth, while it falls relative to NoNoise in NoiseFS. There are strong treatment effects, although somewhat fewer of the pairwise treatment differences are statistically significant according to group level non-parametric tests than in the binary treatments. On the other hand, the difference between NoNoise and NoiseBoth, which in the binary treatments is significant only in Phase 6, is significant in both phases 6 and 7 in the continuous treatments, according to these tests. This strengthens the view that imperfect information increases the relative popularity of FS. In the continuous treatments, FS is almost as popular in NoiseBoth as in NoiseIS. In the binary treatments, FS is most popular in the NoiseIS treatment, although the difference from NoiseBoth is not statistically significant.

Contributions and earnings

To begin to reach a deeper understanding of what lies behind the strong treatment effects on voting, we next investigate treatment effects on contributions to the public good, and on earnings. As mentioned in the Introduction, presence of errors tends to reduce earnings by a substantially larger amount in IS than in FS, and this plays an important part in explaining the asymmetric impact of errors on the choice of institution. Figure 3 shows average contributions to the public good, and average number of punishment points received in IS, by treatment, institution and period of the experiment. Table 4 presents average contributions (Panel A) and earnings (Panel B) by treatment and phase of the experiment. The table also presents tests of the effect of errors in IS and FS, respectively, on contributions and earnings under the different regimes.

[19]Approximately the same conclusion was reached when other, related methods for estimating the share of the treatment effects that work through income were used, for example, when income in earlier periods than the "most recent" are added to model (3), these variables are entirely insignificant, and the estimated share of the treatment effects, which is mediated by income, is still about 50 percent.

Figure 3 Contributions and punishment by treatment, institution and period



Note: NS stands for "no sanction", i.e. the standard voluntary contribution mechanism, which we observe only in Phase 1. N = 1,316 (47 groups observed in 7*4 = 28 periods each).

Figure 3 shows that groups do indeed face collective action problems in the Phase 1 VCM. Although most subjects start out contributing their endowment to the public good, the majority are free riders by period 4. The downward trend in average contributions is visible in all treatments, as expected given that the information and incentives facing subjects are identical.

Table 4 Treatment effects on contributions and earnings

		Groups using IS			Groups using FS		
Panel A: Contributions	Phase 1 (VCM)	Phase 2	Phases 4-7	All observations of groups using IS	Phase 3	Phases 4-7	All observations of groups using FS
NoNoise	12.0	14.3	18.0	16.7	18.9	18.7	18.8
NoiseIS	11.0	9.5	14.3	10.7	19.3	19.3	19.3
NoiseFS	13.4	14.3	18.1	17.1	16.8	18.9	18.0
NoiseBoth	13.3	12.4	13.9	13.2	16.6	17.1	17.0
Noise in IS (NoiseIS + NoiseBoth vs. NoNoise + NoiseFS), p-value		0.02**	0.03**	0.00***			
Noise in FS (NoiseFS + NoiseBoth vs. NoNoise + NoiseIS), p-value					0.00***	0.03**	0.00***
Panel B: Earnings							
NoNoise	32.0	28.5	36.7	33.9	35.0	34.7	34.8
NoiseIS	31.0	24.1	28.9	25.3	35.7	35.8	35.7
NoiseFS	33.4	27.5	35.5	33.4	30.1	33.7	32.1
NoiseBoth	33.3	26.4	29.3	27.9	29.9	30.7	30.5
Noise in IS (NoiseIS + NoiseBoth vs. NoNoise + NoiseFS), p-value		0.23	0.01***	0.02**			
Noise in FS (NoiseFS + NoiseBoth vs. NoNoise + NoiseIS), p-value					0.00***	0.00***	0.00***

Entries in the top four rows of each panel are average contributions/earnings per period per person. The tests in the "Noise in IS" and "Noise in FS" rows are based on linear random effects group level regressions including only dummies for noise in IS/noise in FS, respectively (standard errors clustered at group level). N = 1,316 (47 groups observed in 7*4 = 28 periods each).

Beginning with Phase 2, the situation facing groups using IS differs in the NoNoise and NoiseFS treatments from that in the NoiseIS and NoiseBoth treatments. Considering first separately and then together play of IS when exogenously assigned (Phase 2) and when chosen by vote over FS (relevant observations from Phases 4 – 7), Panel A of Table 4 shows that average contributions are always lower when observational errors are present than when they are not. This difference is statistically significant both in Phase 2 and in Phases 4 - 7, but much stronger in the latter case than in the former. Interpretively, the effect of noise on contributions in IS is growing over time, although its observation only in groups voluntarily selecting it in Phases 4 - 7 should if anything reduce the size of that effect.[20] Panel B of Table 4 shows that earnings in IS

[20] We would expect, that is, that groups that suffer less behavioral impact from errors in IS tend to vote for the system more than do groups with larger impact, all else being equal.

are also always lower when errors are present than when they are not, although this difference is only statistically significant in phases 4 - 7.

Turning to groups operating with either exogenous (Phase 3) or endogenous (Phases 4 – 7) FS, a similar pattern as in the groups using IS emerges. Contributions as well as earnings are lower when observational errors are possible than when they are not (note that earnings in FS depends more or less deterministically on contributions, contrary to what is the case in IS).[21] These differences are statistically significant in both phase 3 (exogenous FS) and phases 4 – 7 (endogenous FS) for both contributions and earnings. In contrast with IS, however, the effects are smaller in the endogenous than in the exogenous parts of the experiment. One interpretation is that in IS, it takes some time to learn how to use the institution effectively, and that this learning process only works well when there is no noise (earnings under IS increase much more strongly from phase 2 to phases 4 - 7 when there is no noise than when there is noise). In FS, on the other hand, subjects may initially be uncertain about what the optimal strategy is when information is noisy, but gradually learn that contributing to the public good is the income maximizing choice.

Comparing earnings in phase 1 (VCM) with earnings in phases 4 – 7, it is interesting to see that for both IS and FS, earnings increase over time in the treatments without noise and decrease in the treatments with noise. Overall, for both IS and FS, sanctioning institutions increase efficiency when information is perfect but not when it is noisy. The differences in earnings versus NS are significant for the decrease under IS with noise ($p = .01$) and for the increase under FS without noise ($p < .01$) and insignificant for the increase under IS without noise ($p = .60$) and the decrease under FS with noise ($p = .31$), Wilcoxon signed-rank tests. In FS, the increase in earnings in treatments without noise is present already in the phase with exogenous institutions, whereas in IS earnings are actually lower in the exogenous phase (phase 2) than in phase 1. Again, some learning is required for IS to be used effectively, even when information is perfect.

An interesting observation emerges from focusing on the "NoiseFS" treatment in Panel B of Table 4. In this treatment, earnings with noisy FS in Phase 3 (30.1 points) were higher than in noise-free IS in Phase 2 (27.5 points). Nevertheless, Figure 2 shows that 60 percent of groups voted for IS in Phase 4 (70 percent on average in phases 4-7). Possibly, participants correctly anticipated an upward trend in earnings under IS. However, it is also possible that some subjects were repelled by an institution in which unjust punishment occurs with predictable frequency, and thus voted for (error free) IS due to normative reasons.

[21] Randomness with respect to which periods experience errors when errors in FS are possible plays little role at the level of analysis undertaken by us, since error frequency tends to converge towards the 10% expected level as the number of observations becomes large. The non-deterministic component present for IS but not FS is that individual group members decide whether or not to punish, in IS, and their decisions may also be affected by knowing that reports are erroneous with likelihood 0.1.

In sum, results show significant negative effects of noise on contributions and earnings in both IS and FS, consistent with the results reported in Ambrus and Greiner (2012) and Grechenig, Nicklisch and Thöni (2010) for IS and in Dickson, Gordon and Huber (2009) and Markussen, Putterman and Tyran (2016) for FS. The fact that noise in either treatment reduces earnings contributes (as also attested by Table 2's regression results) to explaining the strong treatment effects on voting for IS vs. FS, especially the considerable "flight" from the institution afflicted with errors in the two asymmetric treatments (NoiseFS and NoiseIS) relative to NoNoise. The effects of noise on earnings is stronger in IS (a 20 percent drop) than in FS (a 12 percent drop). This contributes to explaining the shift in voting toward support for FS in the NoiseBoth treatment, as compared to the NoNoise treatment (see Figure 2), as well as the greater flight from IS in NoiseIS versus flight towards IS in NoiseFS.[22]

The results are not in line with standard economic theory, which predicted no effects of noise on contributions (cf. Section 4). The results for IS, however, are well in line with the behavioral model of inequality aversion presented in Section 4, which predicts that cooperative equilibria are more rare when information is noisy than when it is perfect ("Factor 2" in the framework presented in Section 4). To be sure, an objection to this view might be raised by noting that the model fails to predict the occurrence of actual punishment in IS, punishment that Figure 3 shows did take place. A response could be that the model is parsimonious in assuming that people know the distribution of social preferences in their group and therefore adjust to equilibrium outcomes immediately. Since in fact subjects only have imperfect information about the preferences of their peers, some cost, in the form of punishment, must be incurred in the process of adjusting to an equilibrium (a process which may or may not be concluded within the limited amount of time available in the experiment). As for FS, the effect of errors on contributions in this environment is not predicted by either standard theory or by the behavioral models we have presented, but contributions in FS with and without noise, respectively, do appear to be converging over time. A possible explanation for the effect of noise on contributions in FS is that some subjects found it difficult to calculate the income maximizing strategy in FS with noise, but many show signs of learning.

Results on earnings and contributions in the treatments with continuous contribution decisions are presented in Figure C2 and Table C4 in Appendix C. Results are again similar to those from the treatments

[22] We refrain from testing for the statistical significance of differences in contributions and earnings between IS and FS of a given phase and treatment due to the small numbers of observations of one or the other regime and the difficulty of interpretation due to endogeneity of regime, for example observations of IS when unpopular in most groups may be mainly those of groups that operate unusually well in that regime. For what it is worth, regression-based tests comparing all group level observations of groups using FS to corresponding observations of groups using IS find statistically significant differences for contributions only in the NoiseIS and NoiseBoth treatments, and for earnings only in the NoiseIS treatment.

with binary contributions. One minor difference is that in the continuous treatments, the detrimental effect of noise on earnings in IS in phases 4-7 is significantly stronger in the NoiseBoth treatment than in NoiseIS. In the binary treatments, earnings under IS are about equally low in both NoiseIS and NoiseBoth.

Punishment in IS

To further understand the treatment effects of noise on contributions and earnings, this section investigates the effects of noisy information on punishment. In FS, since punishment is administered automatically by the computer, these effects are deterministic, given the random occurrence of information errors. In IS, on the other hand, punishment is administered by subjects, who choose themselves how to respond to the information they receive. We therefore focus on punishment in IS.

Table 5 presents descriptive statistics on targeting of punishment in IS. The results clearly demonstrate that noisy information renders targeting less precise. In the treatments without noise in IS, 78 percent of free riders and 12 percent of contributors receive at least one punishment point in periods in which IS is in place (a "free rider" is here a contribution of zero in a given period, so the same person can appear in this analysis as a free rider in some periods and as a contributor in others).[23] In IS with noise, on other hand, 72 percent of free riders and 22 percent of contributors receive sanctions (Panel A)—classifying by actual as opposed to reported behavior. Panel B shows the mean number of punishment points given. The results show that free riders in treatments without noise in IS on average were given three punishment points, while contributors were given 0.2 points. Since one punishment point directed at the subject reduces earnings by four points, punishment reduces the expected earnings of free riders in IS without errors by (3 – 0.2)*4 = 11.2 points, relative to contributors. Since the first-order gain from free riding, rather than contributing, is (1 – a)*W = 0.6*20 = 12 points, this is close to being "behaviorally deterrent", i.e. close to cancelling the pay-off gain from free riding.

[23] The presence of some punishing of contributors even in the absence of noise is consistent with the findings of perverse or antisocial punishment in most contribution and punishment experiments (Cinyabuguma et al., 2006; Herrmann et al., 2008).

Table 5 Targeting of punishment in IS, by treatment

Panel A: Percent punished

	IS without noise (NoNoise and NoiseFS)	IS with noise (NoiseIS and NoiseBoth)
Share of free riders punished	77.7	71.5
Share of contributors punished	11.8	21.8

Panel B: Mean number of punishment points given to..

	IS without noise (NoNoise and NoiseFS)	IS with noise (NoiseIS and NoiseBoth)
Free riders	3.0	2.1
Contributors	0.2	0.4

Note: Calculations based on data from all observations of groups using IS. Classification as "free riders" and "Contributors" is based on actual contributions, not on the contributions displayed to others.

In the treatments with noise in IS, on the other hand, punishment only reduces the earnings of free riders, relative to contributors, by (2.1 – 0.4)*4 = 6.8 points, well short of the deterrent level of punishment. The results in Table 5 clearly show that factor (1) in the theoretical framework of Section 4—"perverse punishment" rises with errors—plays a role (recall the rise in frequency of contributors being punished, from 0.12 to 0.22, roughly an 83% increase).

Table 6 explores whether factor (2) (hesitation to punish apparent free riders) can also be detected in the patterns of punishment. The table presents dyad level regressions for the number of punishment points received (i.e. the unit of observation is punisher-punishee pairs). Explanatory variables include dummies for the joint contribution profiles of senders and receivers of punishment, e.g. "sender contributed, receiver did not contribute", with "sender contributed, receiver contributed" as the reference category. Note that these categorizations are based on actual behaviors of senders and *reported* behaviors of receivers, with the latter failing to match actual behaviors in about 10 percent of cases in NoiseIS and NoiseBoth. A dummy for noise in IS (i.e. for being in NoiseIS or NoiseBoth) is also included and some specifications include the interactions between "Noise in IS" and the contribution indicators.

Table 6 Determinants of punishment in IS

	Dependent variable: Punishment points given to receiver					
	All	Phase 2	Phases 4-7	All	Phase 2	Phases 4-7
	(1)	(2)	(3)	(4)	(5)	(6)
Ref. cat: Sender contributed, receiver contributed						
Sender contributed, receiver did not contribute	0.886***	0.746***	1.044***	1.066***	0.937***	1.191***
	(0.079)	(0.079)	(0.134)	(0.137)	(0.113)	(0.233)
Sender did not contribute, receiver contributed	0.201***	0.223***	0.137*	0.295***	0.337***	0.184*
	(0.051)	(0.055)	(0.071)	(0.078)	(0.098)	(0.103)
Sender did not contribute, receiver did not contribute	0.396***	0.469***	0.246***	0.357***	0.405***	0.212***
	(0.056)	(0.059)	(0.057)	(0.071)	(0.078)	(0.077)
Noise in IS	-0.083*	-0.139***	0.010	0.026	-0.026	0.071**
	(0.044)	(0.045)	(0.062)	(0.027)	(0.027)	(0.029)
Noise in IS * (Sender contributed, receiver did not contribute)				-0.368**	-0.358**	-0.320
				(0.153)	(0.150)	(0.250)
Noise in IS * (Sender did not contribute, receiver contributed)				-0.222**	-0.234**	-0.131
				(0.090)	(0.112)	(0.117)
Noise in IS * (Sender did not contribute, receiver did not contribute)				0.003	0.041	0.030
				(0.100)	(0.112)	(0.097)
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
N	9,520	3,760	5,760	9,520	3,760	5,760

Note: Random effects linear regressions. Units of observation are sender-receiver pairs (dyads). Sender-receiver classifications according to actual sender behavior and reported receiver behaviors. Standard errors clustered by group.

Results show, as expected, that punishment is most severe when the sender contributed and the receiver was reported as a free rider (we refer to this as "pro-social punishment", cf. Herrmann et al., 2008, Fu and Putterman, 2017). However, there is also a significant amount of perverse and antisocial punishment, i.e. cases where the *sender* was a free rider while the *receiver* contributed, according to the public signal.[24] The dummy for noise in IS is significant only in phase 2 (exogenous IS), not in phases 4-7 (endogenous IS). This shows that in phase 2, the possibility of error led to significantly lower punishment, indicating that hesitation to punish (factor (2) of the theoretical framework) does in fact play a role. This result contrasts somewhat with the results in Ambrus and Greiner (2012) and in Grechenig, Nicklisch and Thöni (2010). The latter paper, in particular, emphasizes the finding of "punishment despite reasonable doubt." Regressions 4 and 5 show that the negative effect of noise in IS is mostly driven by a decrease in pro-social punishment, as predicted by the theoretical model. Remarkably, however, there is also a significant reduction of

[24] Cases of punishment when the sender contributed and the receiver was also reported as a contributor also count as both perverse and antisocial punishment, according to the sources just cited. While Table 6 shows that punishment in such dyads (the reference category) was less common than in other types of dyads, it did occur in about 2 percent of the possible cases.

perverse and antisocial punishment. It appears that just as contributors are averse to the possibility of punishing a fellow contributor, free riders are also loath to risk punishing a fellow free rider!

So, not only do subjects receive less precise information in the treatments with noise; the results of this analysis show that they also respond less strongly to that information. This could result from aversion to advantageous inequality, as suggested by the theoretical model. As highlighted in the same model, in order to maintain a deterrent level of punishment, punishers need to respond *more strongly* to reported contributions when information is noisy than when it is not. Hence, these results contribute to explaining why punishment is less effective in IS with noise, and therefore to explaining why noisy information undermines the popularity of IS, relative to FS.[25]

The fact that noise reduces peer punishment may partly reflect the same normative or preference element, as opposed to simple payoff-seeking motivation, that we conjecture is at work in the bias towards FS over IS when both suffer observational error with the same probability. Some subjects, that is, are less ready to punish an apparent free rider when there is the possibility they are thereby unjustly punishing a cooperator. Both dislike of carrying out, and dislike of receiving, unjust punishment at the hands of other individuals, may incline subjects towards FS with noise as the "lesser of two evils."

Results from the treatments with continuous contribution decisions are presented in Tables C5 and C6 in Appendix C. They are again qualitatively similar to results from the binary treatments. In particular, punishment is less well targeted in the treatments with noise than in those without (Table C5), and the effect of low, reported contributions on the probability of receiving punishment from high contributors is lower in the treatments with noise than in those without, although this effects is only statistically significant in the first part of the experiment, as is the case in the binary treatments (Table C6).

The relative importance of the explanatory factors in explaining the treatment effects on institutional choice

Section 4 pointed to three different factors that may explain why IS with noise is less popular than IS without noise, namely (1) increased punishment of cooperators ("perverse punishment"), (2) reduced cooperation due to "scruples" about punishing those displayed as free riders, given the danger of punishing a cooperator, and (3) betrayal aversion and the psychological cost of "wrestling with scruples." It was noted that factor (1) applies to FS as well as to IS, while factors (2) and (3) do not. Is there a way to assign shares

[25] Table D1 in the appendix presents analyses of the effects of punishment and information errors on contributions, Results show, among other things, that the effects punishing free riders in IS on contributions in the subsequent period is positive, but only statistically significant in treatments without noise in IS.

of responsibility for the stronger effect of noise on the popularity of IS than of FS, among these three factors?

In the discussion of Table 3, we argued that factors other than income, i.e. factor (3), explains about 50 percent of the treatment effects on institutional choice, effects that are statistically significant for NoiseIS and NoiseBoth but not NoiseFS, hence largely representing behavioral bias against IS.

The relative importance of factors (1) and (2), in accounting for the remaining 50 percent, can be estimated by analyzing why the introduction of noise led to a stronger income decline in IS than in FS. In the analysis shown in Appendix Table D2, we conclude that factor (1)—increased punishment of cooperators—led to a stronger income loss in FS than in IS, and therefore contributes negatively (if at all) to explaining why the introduction of noise resulted in a larger income loss in IS than in FS. Hence, only factor (2)—scruples about potentially punishing cooperators—explains the remaining half of the bias against IS, whereas factor (1) if anything should induce a bias towards FS. In accounting terms, we report in that table that factor (1) explains about -30 percent of the difference in the impact of noise on earnings in FS and IS, respectively, while factor (2) explains 130 percent. The results presented in tables 5 and 6 show that punishment was indeed less precisely targeted in IS as a result of noise and that subjects did to some extent hesitate to punish apparent free riders in IS when noise was present. In accounting terms, our analysis suggests that factor (1) explains about -15 percent (0.5 * -30 percent) of the treatment effects on institutional choice, factor (2) about 65 percent (0.5 * 130 percent) and factor (3) about 50 percent.

5. Discussion and Conclusion

Understanding the emergence of efficiency enhancing institutions is a key challenge for economists, including scholars of economic growth, environmental problems, and innovation (cf. e.g. North 1990, Ostrom 1990, Acemoglu and Robinson 2005). Public goods (including a legal system, protection against crime, enforcement of safety regulations, and much else) can be provided by a variety of mechanisms, and much recent experimental research has shown a surprising capacity of decentralized systems to address at least some collective action problems, in part thanks to behavioral propensities such as an inclination to engage in costly punishment of free riders. The results presented in this paper suggest that the information environment is an important determinant of the relative attractiveness of centralized versus decentralized enforcement institutions. We let subjects in a public goods experiment choose between peer-to-peer sanctions and a centralized punishment authority, while varying across treatments the quality of information about contributions to the public good available to peer punishers and to the central authority,

respectively. Even though the amount of noise we introduce is moderate (a 10 percent chance that a contribution is reported incorrectly) and leaves standard theory predictions unaffected, we observe strong effects of variations in the information environment on choice of institutions. Two main findings emerge. First, the institution that comes with the most accurate information is always by far the most popular in our set-up. Second, noisy information undermines the popularity of peer punishment more than it does the popularity of a rule-bound centralized punishment institution. These results are robust to having either binary or quasi-continuous contribution decisions in the public goods (voluntary contribution) game.

Our results are poorly explained by standard economic theory, not only insofar as efficacy of peer punishment and its popularity relative to formal sanctions are not predicted by it in the first place, but also insofar as the differential effects of noise on peer vs. centralized punishment are not anticipated by it. Our behavioral theoretical framework suggests that three factors are important for understanding why informal sanctions are less popular when information is noisy than when it is not. First, observational errors lead to increased incidence of "perverse punishment" (punishment directed toward cooperators). However, the change in such punishment's incidence is if anything even more operative in the formal sanctions environment, where punishment is triggered by the signal, without discretion. This factor is thus unlikely to explain a bias away from informal sanctions when both schemes suffer from equally imperfect information. Second, aversion to advantageous inequality means that some people hesitate to punish apparent free riders when information is noisy. Their "scruples" make it more difficult to obtain a cooperative equilibrium under peer punishment, which would call for more not less punishing when information is imperfect, the upshot being a lowering of earnings when punishment is left to peers. Third, if mis-targeted punishment must occur, people may prefer that such punishment be administered by an impersonal system, and not by individuals (they may suffer from broad-sense "betrayal aversion" insofar as they are the objects of unjust punishment, and disutility from "wrestling with scruples" when considering meting it out). Empirical results suggest that all three factors are indeed important. The effects of the information environment on voting are partly explained by the effects of this environment on contributions, punishment and earnings. These effects may in turn be explained by increased perverse punishment and by hesitation to punish those displayed as free riders in IS. However, the results also suggest that subjects dislike noisy information (and the resulting mis-targeting of punishment) even beyond the effects of information errors on earnings. Broad sense betrayal aversion and dislike of having to punish despite doubt is one potential explanation for their decisions.

One caveat is that the results are found in an environment where the quality of information is "salient". For experimental subjects to understand the rules of the experiment, the quality of information must

necessarily be discussed in the instructions. Although the degree of precision and salience of lab participants' knowledge about information imperfectness may not fully apply in real world environments, however, we believe it is realistic to assume that the quality of information used in decisions on punishment is at least somewhat salient. Issues of wrongful acquittal and, especially, wrongful punishment are often high on the public agenda, as exemplified by attention to the standards for conviction in criminal and civil trials, debates prompted by potential inaccuracy of unmanned speed cameras (CBS News 2013) and grass-roots citizen activism in response to excessive use of force by police.

Our paper is most closely related to the large experimental literature on decentralized sanctioning, and to the recent contributions to that literature which compare decentralized and centralized sanctions to understand what may drive the creation of formal institutions. The findings that peer-to-peer sanctioning can raise contributions (Fehr and Gächter, 2000) and earnings (Gächter, Renner and Sefton, 2008) despite its cost to punishers, that it is often favored over a sanction-free environment, especially with learning (Gurerk, Irlenbusch and Rockenbach, 2006, Ertan, Page and Putterman, 2008), and that it is even popular when moderate cost, perfectly targeted formal sanctions are available (Markussen, Putterman and Tyran, 2014), affirms the view that self-governance is sometimes possible (Ostrom, Walker and Gardner, 1992). But misdirected punishment, present with non-trivial frequency even when information is not a problem (Herrmann, Thöni and Gächter, 2008), and likely to be magnified by opportunities for retaliation (Nikiforakis and Engelmann, 2011) and imperfect information (Grechenig, Nicklisch and Thöni, 2010, Ambrus and Greiner, 2012), can tip the balance in favor of a centralized enforcement regime, even when the authority is not a rule-bound automaton (Fehr and Williams, 2013, Nicklisch, Grechenig and Thöni, 2016). Our results indicate that the effect of the information environment on preferences for peer punishment vs. centralized enforcement may be very strong and that imperfect information may lead to preferences for a centralized regime even when the centralized punisher is as error-prone as are peers. The main reasons may be taste- or behaviorally-based motivations for preferring reliance on an impersonal, centralized authority as the lesser of evils, when punishment-in-error is unavoidable.

One direction for future research that this suggests to us is the desirability of going beyond the treatment of information quality as a given. In our experiment, subjects could choose a governance scheme for managing their problem of collective action, but the frequency of erroneous information was chosen by the experimenter. Many of our subjects showed a disinclination to punish free riders when errors were known to be possible even with the modest probability of 10%. Although the peer sanctioning scheme let them use their judgment about punishing on a case by case basis, they perhaps understood, in the NoiseBoth treatment, that some sanctioning would be beneficial, but voted to give the job to the central authority, to

avoid wrestling with it themselves. Future work should consider the further margin that environments often make available: error probabilities can be reduced at a cost, by monitoring. Which kind of monitoring, centralized or decentralized, is more cost effective, is likewise pertinent and subject to study.

References

Aimone, J., Houser, D. 2013. Harnessing the Benefits if Betrayal Aversion. Journal of Economic Behavior and Organization 89, 1 - 8.

Aimone, J., Houser, D. and Weber, B., 2014. Neural Signatures of Betrayal Aversion: an fMRI Study of Trust. Proceedings of the Royal Society B 281, 20132127.

Ambrus, A. and Greiner, B. (2012). Imperfect Monitoring with Costly Punishment: An Experimental Study. *American Economic Review* 102(7), 3317-3332.

Ambrus, A. and Greiner, B. (2015). Democratic Punishment in Public Goods Games with Perfect and Imperfect Observability. ERID Working Paper no. 183.

Andreoni, J. and Gee, L.K. (2012). Gun for Hire: Delegated Enforcement and Peer Punishment in Public Goods Provision. *Journal of Public Economics* 96(11), 1036-1046.

Aoyagi, M., Bhaskar, V., and , Fréchette, G.R. (2015). The Impact of Monitoring in Infinitely Repeated Games: Perfect, Public, and Private. Working paper.

Berg, J., Dickhaut, J., McCabe, K. 1995. Trust, Reciprocity and Social History. Games and Economic Behavior 10 (1), 122 – 142.

Bohnet, I., Zeckhauser, B. (2004). Trust, Risk and Betrayal. Journal of Economic Behavior and Organization 55, 467 – 484.

Bohnet, I., Greig, F., Herrmann, B., Zeckhauser, R. (2008), Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States, *American Economic Review* 98, 294-310.

Carpenter, J. (2007). Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods. *Games and Economic Behavior* 60(1), 31-52.

Carpenter, J., Kariv, S. and Schotter, A. (2012). Network Architecture, Cooperation and Punishment in Public Goods Experiments. *Review of Economic Design* 16, 93-118.

Cason, T. N. & Khan, F. U. (1999), 'A laboratory study of voluntary public goods provision with imperfect monitoring and communication', Journal of Development Economics 58(2), 533–552.

CBS News (2013). Unmanned speed cameras under fire for accuracy problems. March 14th. (link: http://www.cbsnews.com/news/unmanned-speed-cams-under-fire-for-accuracy-problems/ , accessed on August 9th, 2017)

Chaudhuri, A. (2010). Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature. *Experimental Economics* 14(1), 47-83.

Cinyabugama, M., Page, T. and Putterman, L. (2006). Can Second-Order Punishment Deter Perverse Punishment? *Experimental Economics* 9(3), 265-79.

Diamond, Jared, 2012, The World Until Yesterday. New York: Viking Books.

Dickson, E.S., Gordon, S.C. and Huber, G.A. (2009). Enforcement and Compliance in an Uncertain World: An Experimental Investigation. *Journal of Politics* 71, 1357-1378.

Ertan, E., 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.

Fehr, E. and Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90(4), 980-994.

Fehr, E. and Schmidt, K.M. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of* Economics 114(3), 817-868.

Fehr, E., Williams, T. (2013). Endogenous emergence of institutions to sustain cooperation. Mimeo.

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

Fischer, S., Grechenig, K. and Meier, N. (2016). Monopolizing sanctioning power under noise eliminates perverse punishment but does not increase cooperation. *Frontiers in Behavioral Neuroscience* 10, 1-11.

Fu, T. and Putterman, L. (2017). When is Punishment Harmful to Cooperation? A Note on Antisocial and Perverse Punishment. Unpublished paper, Brown University.

Fu, T., Ji, Y., Kamei, K., Putterman, L. (2017). Punishment can Support Cooperation even when Punishable. *Economics Letters* 154, 84 - 87.

Fudenberg, D., Rand, D.G., Dreber, A. (2012). Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *The American Economic Review* 102(2), 720-749

Gachter, S., Renner, E. and Sefton, M. (2008). The long-benefits of punishment. *Science* 322, 1510

Grechenig, K., Nicklisch, A., and Thöni, C. (2010). Punishment despite Reasonable Doubt – A Public Goods Experiment with Sanctions under Uncertainty. *Journal of Empirical Legal Studies* 7(4), 847-867.

Gross, J., Meder, Z.Z., Okamoto-Barth, S., Riedl, A. (2016). Building the Leviathan - voluntary centralization of punishment power sustains cooperation in humans. *Scientific Reports* 6, 207-267.

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

Hauser, O., Rand, D., Peysakhovich, A. and Nowak, M. (2014). Cooperating with the Future. *Nature* 511, 220 – 223.

Herrmann, B., Thöni, C. and Gächter, S. (2008). Antisocial Punishment Across Societies. *Science* 319(5868), 1362-1367.

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

Kamei, K., Putterman, L. and Tyran, J.-R. (2015), State or Nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods, Experimental Economics 18, 38 – 65.

Kosfeld, M., Okada, A. and Riedl, A. (2009), Institution Formation in Public Goods Games, *American Economic Review*, 99, 1335-1355.

Ledyard, J. (1995), "Public Goods: A Survey of Experimental Research," pp. 111 – 194 in J. Kagel and A. Roth, eds., *Handbook of Experimental Economics*. Princeton: Princeton University Press.

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

Markussen, T., Putterman, L., and Tyran, J.-R. (2014). Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes. *Review of Economic Studies* 81(1), 301-324.

Markussen, T., Putterman, L., and Tyran, J.-R. (2016). Judicial Error and Cooperation. *European Economic Review* 89, 372-388.

Nicklisch, A., Grechenig, K. and Thöni, C., (2016) "Information-sensitive Leviathans," Journal of Public Economics, 144, 1–13..

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. 1990. *Governing the Commons. The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Ostrom, E., Walker, J. and Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86(2), 404-417.

Palfrey, T. R. & Rosenthal, H. (1994), 'Repeated play, cooperation and coordination: An experimental study', The Review of Economic Studies 61(3), 545–565.

Putterman, L., Tyran, J.-R. and Kamei, K.(2011). Public Goods and Voting on Formal Sanction Schemes. *Journal of Public Economics* 96(9-10), 1213-1222

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(1743), 3716-3721.

Yamagishi, T. (1986). The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology* 51(1), 110-116.

Zelmer, J. (2003). Linear Public Goods Experiments: A Meta-analysis. Experimental Economics 6, 299 – 310.

Zhang, B., Li, C., De Silva, H., Bednarik, P., and Sigmund, K. (2014). The Evolution of Sanctioning Institutions: An Experimental Approach to the Social Contract. *Experimental Economics* 51(2), 285-303.

Appendix A: Proofs of theoretical statements.

Appendix B: Instructions from NoiseBoth treatment.

Appendix C: Results from treatments with continuous contribution decisions.

Appendix D: Additional results from treatments with binary contribution decisions.

Appendices

for Online Publication only

Appendix A: Proofs of theoretical statements …… p. 2

Appendix B: Instructions from NoiseBoth …… p. 7

Appendix C: Results from treatments with continuous contribution decision ……. p. 16

Appendix D: Additional results from treatments with binary contribution decisions …. p. 22

Appendix A - Proofs of theoretical statements

Proof that FS is deterrent even with noisy information

(π_i is earnings, E is the expectations operator).

$$\mathrm{E}\left(\pi_i \mid C_i = W\right) = \left(1-x\right)\left(aW + a\sum_{j\neq i} C_j\right) + x\left(aW - sW + a\sum_{j\neq i} C_j\right) - c$$

$$= aW + a\sum_{j\neq i} C_j - xsW - c$$

$$\mathrm{E}\left(\pi_i \mid C_i = 0\right) = \left(1-x\right)\left(W - sW + a\sum_{j\neq i} C_j\right) + x\left(W + a\sum_{j\neq i} C_j\right) - c$$

$$= W + a\sum_{j\neq i} C_j - \left(1-x\right)sW - c$$

$$\mathrm{E}\left(\pi_i \mid \mathrm{C}_i = W\right) - \mathrm{E}\left(\pi_i \mid \mathrm{C}_i = 0\right) > 0$$

$$\Leftrightarrow \left(a-1\right)W - 2xsW + sW > 0$$

$$\Leftrightarrow x < \frac{s+a-1}{2s} = \frac{0.8+0.4-1}{1.6} = 0.125.$$

Proof of Proposition 1

Suppose that one of the players $i \in \left\{n'+1,\ldots,n\right\}$ chooses $C_i = 0$ and that all players stick to the punishment strategies stated in the proposition. Then the earnings of the deviator and of a conditionally cooperative enforcer, *j,* are, respectively:

$$\pi_i = W + a\left(n-1\right)W - \left(1-x\right)n'\frac{W}{n' - \frac{1}{\sigma} - \left(2n' - \left(1+\frac{3-n}{\sigma}\right)\right)x}$$

$$\pi_j = a\left(n-1\right)W - \left(\left(1-x\right) + \left(n-2\right)x\right)\frac{W/\sigma}{n' - \frac{1}{\sigma} - \left(2n' - \left(1+\frac{3-n}{\sigma}\right)\right)x}$$

$$-x\left(n'-1\right)\frac{W}{n' - \frac{1}{\sigma} - \left(2n' - \left(1+\frac{3-n}{\sigma}\right)\right)x}$$

where the second-to-last term in the expression for π_j is the expected cost of punishing those displayed as free riders (taking into account that the deviator is displayed as a contributor with probability *x* and that each of the *n*-2 contributors other than the enforcer herself is displayed as a free rider with probability *x*), and the last term reflects that fact that a conditionally cooperative enforcer may herself be displayed to the others as a free rider, in which case she will be punished by $n'-1$ enforcers (i.e. by all enforcers but herself). The expression for π_j can be rewritten as follows:

$$\pi_j = W + a(n-1)W - \left((1-x)+(n-2)x\right)\frac{W/\sigma}{n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x}$$

$$-x(n'-1)\frac{W}{n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x} - \frac{\left(n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x\right)W}{n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x}$$

$$= W + a(n-1)W - \frac{\left(\frac{1+(n-3)x}{\sigma}+x(n'-1)+\left(n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x\right)\right)W}{n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x}$$

$$= W + a(n-1)W - (1-x)n'\frac{W}{n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x} = \pi_i$$

Hence, a deviator has the same monetary pay-off as an enforcer.

Next , we need to check that a deviator does not earn more than a non-enforcing cooperator. When there is no noise, this is trivially true because a non-enforcing cooperator earns at least as much as an enforcer, since the enforcer pays the cost of enforcement in case there is a deviator. However, when noise is introduced, it might not hold, because cooperators risk being the victims of type I errors. If they are, they are punished by *n'* enforcers, whereas an enforcer who is exposed to a type I error is only punished by *n'*-1 enforcers (i.e. not by herself).

The earnings of a non-enforcing cooperator, *k*, are:

$$\pi_k = a(n-1)W - xn'\frac{W}{n'-\frac{1}{\sigma}-\left(2n'-\left(1+\frac{3-n}{\sigma}\right)\right)x}$$

This is higher than or equal to the earnings of a deviator if:

$$a(n-1)W - xn'\frac{W}{n' - \frac{1}{\sigma} - \left(2n' - \left(1 + \frac{3-n}{\sigma}\right)\right)x} \geq W + a(n-1)W - (1-x)n'\frac{W}{n' - \frac{1}{\sigma} - \left(2n' - \left(1 + \frac{3-n}{\sigma}\right)\right)x}$$

$$\Leftrightarrow \frac{(1-2x)n'}{n' - \frac{1}{\sigma} - \left(2n' - \left(1 + \frac{3-n}{\sigma}\right)\right)x} \geq 1$$

(assumption (6) ensures that the denominator on left-hand side is positive).

$$\Leftrightarrow (1-2x)n' \geq n' - \frac{1}{\sigma} - \left(2n' - \left(1 + \frac{3-n}{\sigma}\right)\right)x$$

$$\Leftrightarrow x \leq \frac{1}{(\sigma + 3 - n)}$$

Assumption (6) ensures that this condition holds, and deviation is therefore not profitable.

Next, we need to show that the threat of punishment is credible. We first calculate the probability assigned by an enforcer to a fellow group member, *j*, being a cooperator, given that *j* is reported as being a free rider. Using Bayes' rule, we get that:

$$b_i(x) = P_i\left(C_j = W \mid C_j^{reported} = 0\right) = \frac{P_i\left(C_j^{reported} = 0 \mid C_j = W\right)P_i\left(C_j = W\right)}{P_i\left(C_j^{reported} = 0\right)} = \frac{xP_i\left(C_j = W\right)}{P_i\left(C_j^{reported} = 0\right)}$$

$$= \frac{xP_i\left(C_j = W\right)}{xP_i\left(C_j = W\right) + (1-x)\left(1 - P_i\left(C_j = W\right)\right)}$$

where $P_i(C_j = W)$ is the enforcer's prior belief about the probability that *j* is a cooperator. It is simple to show that $b_i'(x) > 0$ for $0 < P_i(C_j = W) < 1$, i.e. as long as the enforcer's prior assigns positive probabilities to *j* being a cooperator as well as to him or her being a free rider, her posterior belief that *j* is a cooperator, given that he is reported as a free rider, increases in *x*.[26] It is also clear that $b_i(x) = 0$ when $x = 0$.

Now, assume that an enforcer observes one fellow group member who is displayed as contributing less than *C*. If a punisher reduces p_{ij} by ε, she saves ε and experiences less disadvantageous inequality relative to those $n - n' - 1$ players who contribute to the public good but do not punish. This creates a

[26] Note also that $b_i(x) = x$ for $P_i(C_j = W) = 0.5$. In words, if enforcers initially keep an "open mind" about whether fellow group members are free riders or cooperators, then $b_i(x)$ reduces to x.

nonmonetary utility increase of $\left(\alpha_i\left(n-n'-1\right)\varepsilon\right)/\left(n-1\right)$.[27] The change in utility due to changes in inequality vis a vis the recipient of punishment depends on whether the recipient was correctly displayed as a free rider, or not. If the recipient of punishment was in fact a free rider, the enforcer also has nonmonetary *costs* from reducing p_{ij} because she now experiences disadvantageous inequality toward the free rider, amounting to $\alpha_i\left(\sigma-1\right)\varepsilon/\left(n-1\right)$. On the other hand, if the group member displayed as a free rider is in fact a cooperator, the enforcer will benefit from reduced, advantageous inequality toward the recipient of punishment, generating a utility gain of $\beta_i\left(\sigma-1\right)\varepsilon/\left(n-1\right)$. Hence, the expected change in utility from changes in inequality vis a vis the recipient of punishment is $-\left(\sigma-1\right)\varepsilon\left(\left(1-b_i\left(x\right)\right)\alpha_i-\beta_i b_i\left(x\right)\right)/\left(n-1\right)$. The enforcer experiences a loss due to increased advantageous inequality relative to the $\left(n'-1\right)$ other enforcers who punish fully. This amounts to a utility loss of $\left(\beta_i\left(n'-1\right)\varepsilon\right)/\left(n-1\right)$. Thus, the expected loss from a reduction in p_{ij} is greater than the gain if

$$\frac{1}{n-1}\left[\left(\sigma-1\right)\varepsilon\left(\left(1-b_i\left(x\right)\right)\alpha_i-\beta_i b_i\left(x\right)\right)+\left(\beta_i\left(n'-1\right)\varepsilon\right)\right]>\varepsilon+\alpha_i\left(n-n'-1\right)\frac{\varepsilon}{n-1} \quad \text{(A1)}$$

holds. This condition is equivalent to condition (5). Hence the punishment strategies are credible.

Now, it may happen that an enforcer observes more than one fellow group member, who are displayed as free riders. Will she be willing to punish all of them? In fact, the condition for punishing each free rider is:

$$\frac{1}{n-1}\left[\left(\sigma-1\right)\varepsilon\left(\left(1-b_i\left(x\right)\right)\alpha_i-\beta_i b_i\left(x\right)\right)+\left(\beta_i\left(n'-1\right)\varepsilon\right)\right]>\varepsilon+\alpha_i\left(n-n'-n^{FR}\right)\frac{\varepsilon}{n-1}$$

Where n^{FR} is the number of players displayed as free riders. This is easier to meet than A1, and the enforcer will therefore simply punish everyone displayed as a free rider (the intuition is that if the enforcer reduces punishment, she gains from reduced, disadvantageous inequality toward the non-enforcing contributors. When one group member is moved from the group of non-enforcing contributors to the group of (supposed) free riders, this gain decreases, and punishment becomes more attractive).

We must also check that enforcers have no incentive to deviate in the first stage. If an enforcer free rides, she gains $\left(1-a\right)W$ in earnings but suffers a utility loss of βW due to increased advantageous inequality

[27] It is possible that even if these players were displayed as contributors, they are in fact free riders. However, if they are free riders, they still have higher income than the enforcer, who therefore still reduces disadvantageous inequality toward these players if she reduces punishment directed toward a player displayed as a freer rider.

toward all other group members. Hence, the total welfare gain is $(1-a-\beta)W$. Since all enforcers have $\beta > 1-a$, the deviation does not increase utility. QED.

Appendix B Instructions from NoiseBoth treatment

Welcome

Welcome to our decision-making experiment. Depending on your decisions and the decisions of other participants, you will be able to earn money in addition to the Y_ guaranteed for your participation. Please read the following instructions carefully.

During the experiment you are not allowed to communicate with other participants. If you have a question, raise your hand. One of us will come to answer your question.

During the experiment your earnings will be calculated in points. At the end of the experiment points will be converted to RMB at the following rate:

1 points = 0.055yuan

At the end of the experiment your total earnings (including the RMB10 yuan participation fee) will be paid out to you privately in cash.

The experiment has seven phases each consisting of 4 periods (in total, 28 periods). The following instructions explain the details of phase 1. The details of the subsequent phases will be explained later.

Instructions for Phase 1

In the experiment, each participant is randomly assigned to a **group of 5**. This means that you are in a group with four other participants. **You will be part of the same group throughout the entire experiment**. Nobody knows which other participants are in their group, and nobody will be informed who was in which group after the experiment.

Phase 1 is divided into 4 periods. In each period, each group member, yourself included, will be given an **endowment of 20 points**. In each period you will have to make one decision.

Your decision

You and the four others in your group simultaneously decide how to use the endowment. There are two possibilities:

1. **You can allocate the endowment to a group account.**
2. **You can allocate the endowment to a private account.**

Your earnings depend on the total number of points in the group account, and the number of points in your private account.

How to calculate your earnings

Your earnings from your private account are equal to the number of points you allocate to it (0 or 20). That is, **for each point you allocate to your private account you get 1 point as earnings**. The points you allocate to your private account do not affect the earnings of the others in your group.

Your earnings from the group account equal the **sum** of points allocated to the group account by all 5 group members multiplied by 0.4. **For each point you allocate to the group account you and all others in your group each get 0.4 points as earnings**. For example, if the sum of points in the group account is 40, then your earnings from the group account and the earnings of each of the others in your group from the group account are equal to 16 points.

Your earnings can be calculated with the following formula:

20 – (points you allocated to the group account) + 0.4 * (sum of points allocated by all group members to the group account)

Note that you get 20 points as earnings if you allocate your endowment of 20 points to your private account. If you instead allocate your endowment to the group account, your earnings from the group account increase by 0.4 * 20 = 8 points and your earnings from your private account decrease by 20 points. However, by allocating 20 points to the group account, the earnings of each of the other 4 group members also increase by 8 points. Therefore, the total group earnings increase by 8* 5 = 40 points. Note that you also obtain earnings from points allocated to the group account by others. You obtain 0.4 * 20 = 8 points for each other group member who allocates his or her endowment to the group account.

Example

Suppose you allocate 0 points to the group account, the second and third members of your group each allocate 20 points to the group account, and the remaining two individuals allocate 0 points each. In this case, the sum of points in the group account is 0 + 20 + 20 + 0 + 0 = 40 points. Each group member gets earnings of 0.4 * 40 = 16 points from the group account.

Your total earnings are: 20 – 0 + (0.4 * 40) = 20 + 16 = 36 points.

The second and third members each earn: 20 – 20 + (0.4 * 40) = 0 + 16 = 16 points.

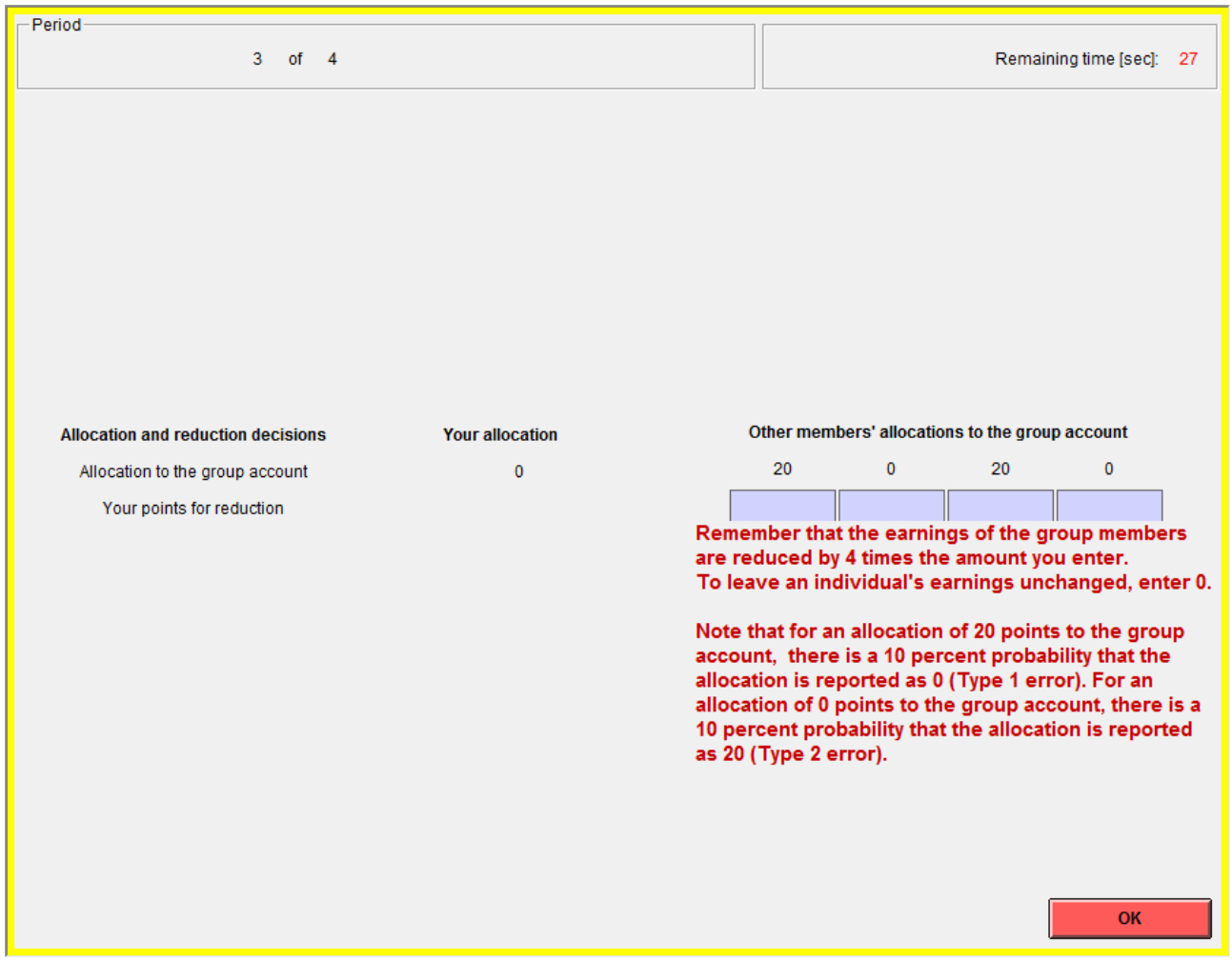
The fourth and fifth members each earn: 20 – 0 + (0.4 * 40) = 20 + 16 = 36 points.

Instructions for Phase 2

The four periods of Phase 2 are like the previous four periods in that you continue to be grouped with the same four individuals and each period begins with an allocation stage having the same consequences for your earnings. However, the periods of Phase 2 include a second stage in which your earnings may be reduced by your own and other group members' decisions. In particular, in the second stage of each period, you have an opportunity to reduce the earnings of others in your group at a cost to your own earnings, and the other group members each have corresponding opportunities. Here is how it will work.

After the first stage of each period, you will be shown the amount allocated to the group account by each of the others in your group, **in a random order**, and in a box below that information on each individual's allocation you will be asked to enter a whole number of points (if any) that you wish to use to reduce the earnings of that individual. Each point you allocate to reducing another's earnings **reduces your own**

earnings by 1 point and **reduces that individual's earnings by 4 points**. Your own earnings can be reduced in the same way by the decisions of others in your group. You are free to leave any or all others' earnings unchanged by entering 0's in the relevant boxes.



Note: Numbers shown are for illustration only.

There are two further details to be aware of.

First, the amounts that other group members allocate to the group account vs. to their private account in a given period will not always be reported accurately. In particular, every time you or another group member allocates 20 points to the group account, with a 10 percent probability there is an erroneous report that the individual allocated 0 to the group account (and hence 20 to the private account)—this is called a **"Type 1" error**. Every time you or another group members allocates 0 points to the group account, there is a 10% probability of an erroneous report that the individual allocated 20 to the group account (and hence 0 to the private account) —this is called a **"Type 2" error**. There will be no way for you to know whether such reports accurately reflect fellow group members' behaviors or instead result from the occurrence of errors with the indicated random probabilities. Your own allocation to the group account is likewise expected to be erroneously reported at the same error rates, i.e. 10% for Type 1 error and 10% for Type 2 error.

Type 1 and Type 2 errors do not affect your *actual* allocation to the group account, and thus do not affect what each group member earns from the group account, which is 0.4 times the total amount of points allocated to it. They also do no not affect what you earn from your private account, which is 1 times what you allocate to the account (or equivalently, 20 minus what you put in the group account). Errors only effect how your allocation is *reported*. The same is true for other group members.

At the end of each period, you will learn whether the report of your allocation was affected by error. However, you will not learn whether the reports about group members' allocations were affected by error. Likewise, other group members will also only learn whether their own allocation was reported correctly or not.

Second, there is an exception to the rule that you lose 4 points for each reduction point another group member assigns to you: it is that reductions cannot bring your earnings for the period to less than zero. So if, for instance, you earned 20 points in the first stage and others assigned a total of 6 reduction points to you, your earnings in the period cannot be lowered by 6x4 = 24, but only by 20. However, the cost of giving reductions to others is always fully born even if it makes your period earnings negative. If you lose points in a period, they are deducted from those you accumulate in other periods. Thus, your earnings in each period of this phase can be calculated as follows:

$$\begin{aligned}\text{Earnings} = [&20 - (\text{points you allocate to group account}) + 0.4*(\text{sum of points allocated by all in group to group account}) - \\ &4*(\text{sum of reduction points directed at you by others in your group})] \\ &- (\text{points you use to reduce others' earnings})\end{aligned}$$

with the exception that the amount in the first three lines (within the brackets [ ]) will be set to zero if it is negative.

For example, suppose that you use 0 points to reduce the earnings of the first and second group members whose allocations appear on the screen, you use 1 point to reduce the earnings of the third, and you use 2 points to reduce the earnings of the fourth. Then the third and fourth individuals' earnings for the period will be reduced by 4 and by 8 points, respectively, in addition to any reductions due to the decisions of others, although these reductions cannot bring their earnings below zero. Suppose further that these individuals use 0, 1, 0 and 3 points to reduce your earnings. Your earnings for the period will be reduced by (1x4)+(3x4)=16 points, as long as that reduction does not bring your period earnings to less than zero. Your earnings will also be reduced by an additional 3 points, your own cost to impose reductions on others. That 3 point cost is incurred by you even if it causes your overall earnings for the period to become negative. At the end of the reduction stage, you will learn that others decided to reduce your earnings by a total of 16

points, but you will not be told which individuals reduced your earnings or by how much any given individual reduced your earnings. Others will also not know who in particular reduced their earnings by how much.

In addition to the fact that earnings from the allocation stage and reductions received cannot go below zero, the earnings reduction process is subject to two limits. First, your reduction points must be an integer. Second, you cannot assign more than 10 reduction points to any one individual in your group.

If no reductions are imposed (the reduction boxes are filled in with 0's), earnings after the reduction stage are the same as those before it. Entering 0 in any or all reduction boxes is always an option.

Instructions for Phase 3

The next four periods are like the previous eight in that you continue to interact with the same four individuals, and in the initial stage of each period you make a decision about allocating 20 points to either a private account or a group account. The earnings consequences of your initial decisions are also as before, and as in Phase 2, there is a second stage of the period in which some of the earnings from the first stage may be lost.

Unlike in Phase 2, however, earnings reductions in this phase are ones that result from the operation of an administrative scheme that functions automatically. Specifically, under the scheme in place during the present phase, if a group member is recorded as putting his or her endowment in their private account, he or she pays a fine of 16 points. Thus, although as in Phase 1 you earn 20 points if you put your endowment in your private account, in this phase you lose 16 points in the 2nd part of the period due to the presence of the administrative scheme, so what you earn from each point you put in your private account after taking the fine into consideration (in other words, your net earnings) is 4 points (= 20 – 16 points). Your earnings and those of other group members from points put in the group account remain as in Phase 1; that is, you and each other group member earn 0.4 points for each point you or they put into the group account.

There is an exception to the rules described above about fines:

As in the previous phase, there is 10% probability of a **Type 1 error** every time you or another group member allocates 20 points to the group account and a 10% probability of a **Type 2 error** every time you or another group member allocates 0 points to the group account (see the instructions for Phase 2 for a detailed description of Type 1 and Type 2 errors).

The automatically imposed fines are based on *reported* allocations, regardless of whether these reports are true or not. Assume that you allocate 0 points to the group account. If there is no error, you pay a fine equal to 0.8*20 = 16 points. If there is a Type 2 error, you pay a fine of 0.8*0 = 0 points. Alternatively, assume that you allocate 20 points to the group account. If there is no error, you pay a fine equal to 0.8*0 = 0 points. If there is a Type 1 error, you pay a fine of 0.8*20 = 16 points.

Type 1 and Type 2 errors do not affect your *actual* allocation to the group account, and thus do not affect what each group member earns from the group account, which is 0.4 times the total amount of points allocated to it. They also do no not affect what you earn from your private account, which is 1 times what you allocate to the account (or equivalently, 20 minus what you put in the group account). Errors only effect how your allocation is *reported*, and through this may affect the amount of fine assessed. The same is true for other group members.

At the end of each period, you will learn whether the report of your allocation was affected by error. However, you will not learn whether the reports about other group members' allocations were affected by error. Likewise, other group members will also only learn whether their own allocation was reported correctly or not.

Recall that in Phase 2, as well as reducing the earnings of the individual receiving the reduction, reducing earnings in a period's 2nd stage cost something to the group member who chose to impose the reduction. Although the administrative scheme used in Phase 3 operates automatically, there is also a cost to its operation. Specifically, during each period of Phase 3 and in any later periods of the experiment that this administrative scheme might be in place, 3 points of earnings are automatically deducted from each group member's earnings as a cost of operating the scheme.

If there are no errors (80 percent probability), your earnings can be calculated as follows:

Earnings = 20 – (points you allocate to group account) + 0.4*(sum of points allocated by all in group to group account) – 0.8*(points you allocate to private account) – 3

If there is **type 1 error** (10 percent probability), your earnings can be calculated as follows:

Earnings = 20 – (points you allocate to group account) + 0.4*(sum of points allocated by all in group to group account) – 0.8*20 – 3

= 20 – (points you allocate to group account) + 0.4*(sum of points allocated by all in group to group account) – 16 – 3

If there is **type 2 error** (10 percent probability), your earnings can be calculated as follows:

Earnings = 20 – (points you allocate to group account) + 0.4*(sum of points allocated by all in group to group account) – 0.8*0 – 2

Examples

Assume that all group members allocate 20 points to the group account (100 points in total). If your allocation is reported without error, your earnings for the period are 20 – 20 + 0.4*100 – 0.8*0 – 2 = 38 points. You earn 40 points in the allocation stage, but have 2 points deducted as your contribution to the cost of operating the administrative scheme. If there is a type 1 error in the report of your allocation, your earnings are 20 – 20 + 0.4*100 – 16 – 2 = 22 points. If there is a type 2 error, your earnings are 20 – 20 + 0.4*100 – 2 = 38 points.

As another example, assume that each allocates 0 points to the group account. If your allocation is reported without error, your earnings for the period are 20 – 0 + 0.4*0 – 0.8*20 – 2 = 2 points. You earn 20 points in the allocation stage, but pay a fine of 16 points and have 2 points deducted as your contribution to the cost of operating the administrative scheme. If there is a type 1 error in the report of your allocation, your earnings are 20 – 0 + 0.4*0 – 16 – 2 = 2 points. If there is a type 2 error, your earnings are 20 – 0 + 0.4*0 – 2 = 18 points. In that case, you were not assessed a fine for putting 20 points in your private account because you were erroneously recorded as allocating 0 to that account.

Instructions for Phases 4 - 7

The next four phases of the experiment, which will also be the last ones, will resemble either Phase 2 or Phase 3, depending on which of the two arrangements your group chooses to follow in a given phase. You will remain in a group with the same four others and will make four allocation decisions in each phase, like before. At the beginning of each phase, you will vote on whether to use the scheme in which individual group members can reduce others' earnings after learning of their allocations (which will be referred to on the voting screen as "SCHEME 1: individual reduction decisions", i.e. the scheme used in Phase 2) or the scheme in which there is an automatic reduction for each point assigned to your private account (called "SCHEME 2: automatically administered reductions", i.e. the scheme used in Phase 3). Whichever scheme gets the most votes will be in effect for four periods. You can select whichever of the two schemes you

prefer in a given phase (4 periods = a phase) regardless of which scheme your group selected for the previous phase or phases.

The way to calculate your earnings is exactly the same as in Phases 2 or 3, depending on which scheme your group chooses.

In both Scheme 1 (individual reduction decisions) and Scheme 2 (automatically administered reductions), there is a 10 percent probability of Type 1 error and a 10 percent probability of Type 2 error each time an allocation to the group account is reported. As in Phase 3, the automatically imposed fines in Scheme 2 are based on *reported* allocations, regardless of whether these reports are true or not.

There will be no additional instructions. Phases 5, 6 and 7 will each begin with votes like that of phase 4, after the four periods of each preceding phase.

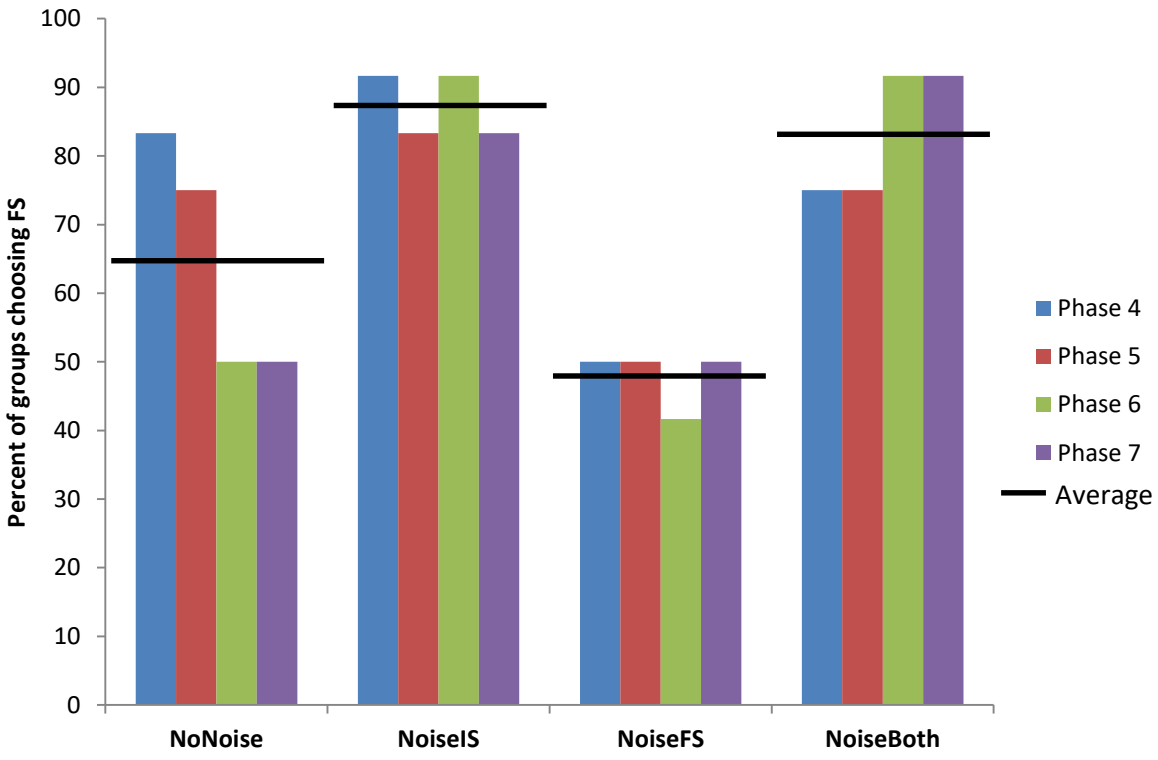
Appendix C Results from treatments with *continuous* contribution decisions

Table C1 Treatments

		Imperfect information in IS	
		No	*Yes*
Imperfect information in FS	*No*	NoNoise (60/12)	NoiseIS (60/12)
	Yes	NoiseFS (60/12)	NoiseBoth (60/12)

Note: Numbers of subjects/number of groups in parentheses. There were two sessions in each treatment. Total number of subjects: 240.

Figure C1 Choice of institution by treatment and phase, group level



Note: N = 48 groups, observed four times each.

Table C2 Mann-Whitney tests of treatment effects on voting, group level

	Phase 4	Phase 5	Phase 6	Phase 7	Mean of voting in all phases	Observations
NoNoise vs. NoiseIS	0.55	0.62	0.03**	0.09*	.09*	24
NoNoise vs. NoiseFS	0.09*	0.22	0.69	1.00	0.44	24
NoNoise vs. NoiseBoth	0.62	1.00	0.03**	0.03**	0.23	24
NoiseIS vs. NoiseFS	0.03**	0.09*	0.01**	0.09*	0.05*	24
NoiseIS vs. NoiseBoth	0.28	0.62	1.00	0.55	0.43	24
NoiseFS vs. NoiseBoth	0.22	0.22	0.01**	0.03**	0.10	24

Note: M-W tests. Entries are p-values.

Table C3 Voting regressions, group level

	Dependent variable: Group voted for FS		
NoiseIS	0.229	0.197	0.127
	(0.137)*	(0.117)*	(0.083)
NoiseFS	-0.167	0.133	0.082
	(0.180)	(0.168)	(0.107)
NoiseBoth	0.187	0.336	0.182
	(0.130)	(0.136)**	(0.084)**
Earnings in Phase 2 (IS)		-0.013	
		(0.002)***	
Earnings in Phase 3 (FS)		0.012	
		(0.006)*	
Earnings in most recent phase with IS			-0.01
			(0.001)***
Earnings in most recent phase with FS			0.007
			(0.003)**
Constant	0.646	0.25	0.729
	(0.110)***	(0.813)	(0.501)
R-sq (overall)	0.12	0.37	0.53
N	192	192	192

Note: Random effects, linear regressions. Standard Noises clustered at group level. The reference treatment is NoNoise.

Figure C2 Contributions and punishment by treatment, institution and period

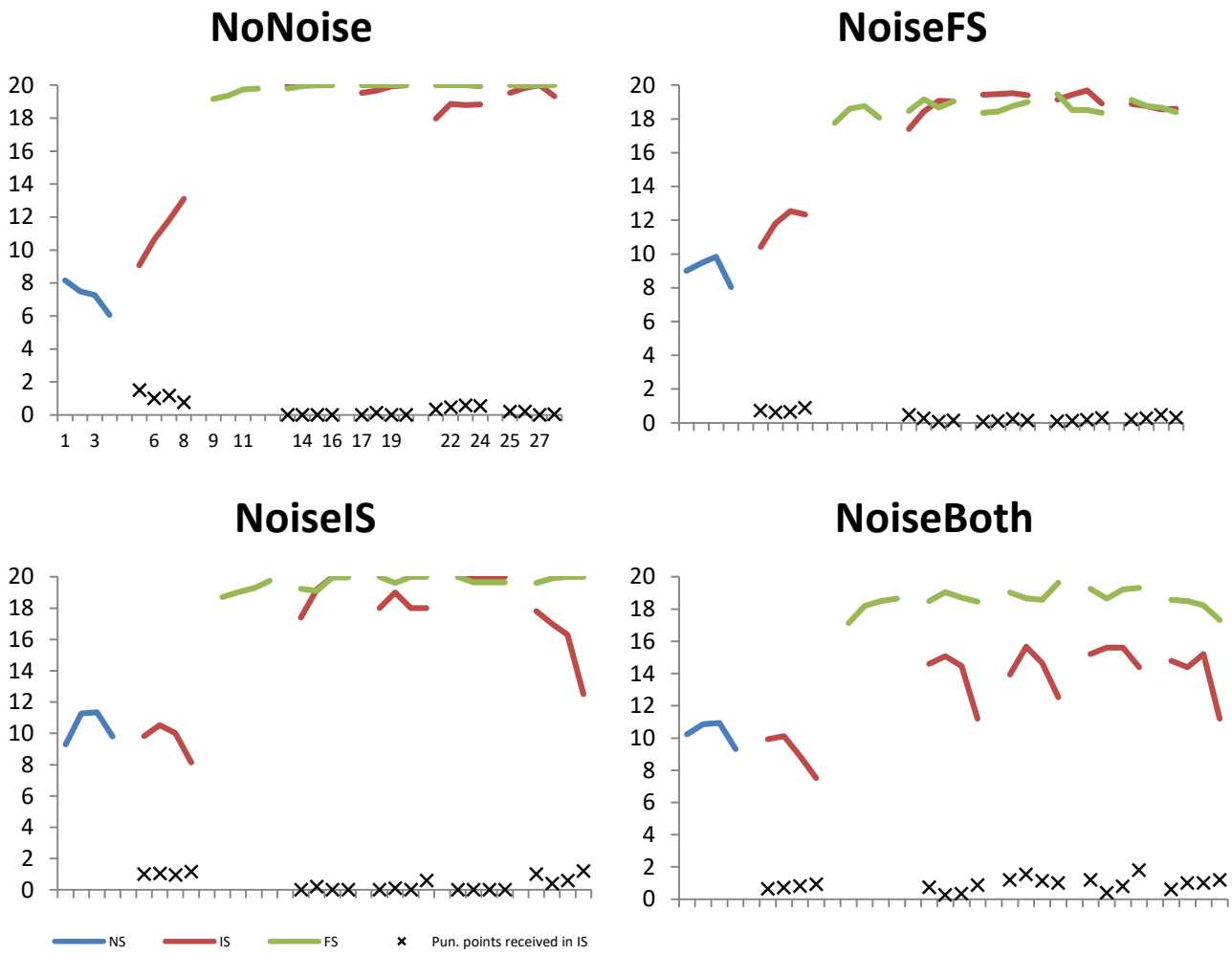


Table C4 Treatment effects on contributions and earnings

		Groups using IS			Groups using FS		
Panel A: Contributions	Phase 1 (VCM)	Phase 2 only	Phases 4-7	All observations of groups using IS	Phase 3 only	Phases 4-7	All observations of groups using FS
NoNoise	7.3	11.2	19.4	16.0	19.5	20.0	19.8
NoiseIS	10.4	9.6	17.9	12.4	19.2	19.8	19.6
NoiseFS	9.1	11.8	19.0	16.7	18.3	18.7	18.6
NoiseBoth	10.3	9.1	14.2	11.1	18.1	18.7	18.6
Noise in IS (NoiseIS + NoiseBoth vs. NoNoise + NoiseFS), p-value		0.04**	0.00***	0.00***			
Noise in FS (NoiseFS + NoiseBoth vs. NoNoise + NoiseIS), p-value					0.00***	0.00***	0.00***
Panel B: Earnings							
NoNoise	27.3	25.6	38.3	33.1	36.1	37.0	36.7
NoiseIS	30.4	24.4	36.2	28.4	35.5	36.6	36.3
NoiseFS	29.1	28.2	38.0	34.8	32.5	33.0	32.8
NoiseBoth	30.3	25.2	29.6	27.0	32.5	33.4	33.2
Noise in IS (NoiseIS + NoiseBoth vs. NoNoise + NoiseFS), p-value		0.18	0.01***	0.00***			
Noise in FS (NoiseFS + NoiseBoth vs. NoNoise + NoiseIS), p-value					0.00***	0.00***	0.00***

Entries in the top four rows of each panel are average contributions/earnings per period per person. The tests in the "Noise in IS" and "Noise in FS" rows are based on linear random effects group level regressions including only dummies for noise in IS/noise in FS, respectively.

Table C5 Targeting of punishment in IS, by treatment

Panel A: Percent punished

	IS without noise (NoNoise and NoiseFS)	IS with noise (NoiseIS and NoiseBoth)
Share of free riders punished	75.0	67.3
Share of contributors punished	11.7	32.2

Panel B: Mean number of punishment points received

	IS without noise (NoNoise and NoiseFS)	IS with noise (NoiseIS and NoiseBoth)
Free riders	2.3	1.7
Contributors	0.2	0.5

Note: Free riders (contributors) are defined as those contributing less than (more than or equal to) the group median contribution in a given period. Observations are classified on the basis of true rather than reported contributions.

Table C6 Dyadic punishment regressions

	Dependent variable: Punishment points given to receiver					
	All	Phase 2	Phases 4-7	All	Phase 2	Phases 4-7
Ref. Cat: Sender contribution >= median, receiver contribution >= median						
Sender contribution >= median, receiver contribution < median	0.441***	0.408***	0.430***	0.563***	0.526***	0.523***
	[0.055]	[0.046]	[0.091]	[0.085]	[0.075]	[0.136]
Sender contribution < median, receiver contribution >= median	0.041*	0.035	0.004	0.073**	0.042	0.033
	[0.022]	[0.022]	[0.035]	[0.029]	[0.034]	[0.041]
Sender contribution < median, receiver contribution < median	0.241***	0.225***	0.225***	0.256***	0.233***	0.198***
	[0.038]	[0.045]	[0.051]	[0.047]	[0.057]	[0.046]
Error in IS	0.011	-0.023	0.032	0.078**	0.032	0.069*
	[0.046]	[0.054]	[0.058]	[0.038]	[0.044]	[0.035]
Error in IS * (Sender contribution >= median, receiver contribution < median)				-0.237**	-0.216**	-0.173
				[0.099]	[0.089]	[0.163]
Error in IS * (Sender contribution < median, receiver contribution >= median)				-0.079*	-0.016	-0.072
				[0.041]	[0.045]	[0.069]
Error in IS*(Sender contribution < median, receiver contribution < median)				-0.043	-0.017	0.023
				[0.069]	[0.086]	[0.092]
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
N	8,320	3,840	4,480	8,320	3,840	4,480

Note: Random effects linear regressions. Units of observation are sender-receiver pairs (dyads). Standard errors clustered by group. Sender-receiver classifications according to actual sender behavior and reported receiver behaviors. The "median" referred to in variable labels is the median contribution in a given group in a given period.

Appendix D - Additional results from treatments with *binary* contribution decisions

Effects of punishment and errors on individual contribution decisions

This section investigates the effects of receiving punishment, and of exposure to observational errors, on individual contribution decisions. Table D1 presents contribution regressions for both IS and FS and splits the sample by a) whether the subject contributed to the public good in the previous period and b) whether noise is present or not. In both IS and FS, the models for treatments with noise include dummies for type I and type II observational errors in the previous period. Note that only free riders can be exposed to type II errors, and only contributors can be hit by type I errors. In IS, we consider the effects of receiving punishment in the previous period, interacted with the error dummies in the treatments with noise. Note that in FS, punishment follows directly from contribution decisions and exposure to errors, and that additional measures of punishment received are therefore redundant. Models also include the average

Table D1 Individual contribution regressions, by institution, treatment and contribution in previous period

	Dependent variable: Contribution to public good ($C_{i,t}$)							
	IS				FS			
	$C_{i,t-1} = 0$	$C_{i,t-1} = 0$	$C_{i,t-1} = 20$	$C_{i,t-1} = 20$	$C_{i,t-1} = 0$	$C_{i,t-1} = 0$	$C_{i,t-1} = 20$	$C_{i,t-1} = 20$
	No noise in IS	Noise in IS	No noise in IS	Noise in IS	No noise in FS	Noise in FS	No noise in FS	Noise in FS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Others' reported mean contribution$_{t-1}$	0.385**	0.131	0.43***	0.469***	-0.884**	0.524***	0.273	0.06
	(0.193)	(0.138)	(0.081)	(0.085)	(0.364)	(0.150)	(0.066)***	(0.058)
Punishment points received$_{t-1}$	0.529*	0.409	-0.285	-1.810**				
	(0.283)	(0.303)	(0.424)	(0.796)				
Type II error$_{t-1}$		0.292				-2.326		
		(3.193)				(2.737)		
Punishment points received*Type II error$_{t-1}$		-1.683						
		(1.480)						
Type I error$_{t-1}$				-1.577				0.841
				(2.131)				(0.514)
Punishment points received*Type I error$_{t-1}$				1.985*				
				(1.087)				
Phase and period FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	169	215	986	415	59	125	1,291	970

Note: Random effects, linear regressions. Constant included (not shown). Level of observation: subjects. All observations of groups using IS/FS included. Standard errors, clustered at group level, in parentheses. Period 1 of each phase is excluded. Samples are defined by the subjects' *true*, rather than *reported*, contributions.

reported contribution of fellow group members in the previous period.

In IS, the results show that punishment of free riders had a positive effect, significant at the 10 percent level, on contributions in the following period in the noise-free environment, and a slightly weaker,

insignificant, but still positive effect when noise was present (models 1 and 2). Punishment of cooperators had a negative effect in both treatments with and without noise, but the effect is only significant when noise is present. Notably, the effect of punishing cooperators is different when punishment follows a type I observational error (i.e. when the cooperator was displayed as a free rider) than when it does not (i.e. when punishment is knowingly directed toward people who most likely cooperated). The numerical effect of "punishment points received*type I error$_{t-1}$" is of the same magnitude, but with opposite sign, as the effect of "punishment points received$_{t-1}$" in model 4, implying that whereas punishment of cooperators displayed as cooperators discourages future contributions (as found in previous papers on perverse/antisocial punishment), punishment of cooperators displayed as free riders does not. One interpretation is that the former type of punishment leads the recipients of punishment to update beliefs about the presence of group members with antisocial preferences, whereas the latter does not. The main effects of exposure to errors are insignificant in all models. This is not surprising, given that subjects knew that errors happened randomly, and were aware of the risk of errors.[28] In sum, punishment does appear to have a significant deterrent effect on free riding in IS without noise, but an insignificant effect in treatments with noise. Punishment of cooperators discourages future cooperation, but this effect is cancelled when the cooperator was displayed to others as a free rider.

The relative importance of the explanatory factors in explaining the treatment effects on institutional choice

Table D2 decomposes the differential impact of noise on earnings in FS and IS (cf. the discussion in the main text about the relative importance of the explanatory factors in explaining the treatment effects on institutional choice).

[28] The most surprising result is the negative effect of others' mean contribution in the previous period in the model for free riders in FS without noise (model 5). The result is consistent with the view that free riding in FS without noise is caused by confusion. Note model 5 includes only 59 observations, from 23 different individuals.

Table D2 Decomposition of changes in earnings

	IS		FS		Difference (IS - FS)		Difference in differences
Panel A	No noise	Noise	No noise	Noise	No noise	Noise	
Total earnings per person per period (points)	33.6	26.9	35.4	31.1	-1.8	-4.1	2.3 (A)
Earnings *before* punishment	36.9	32.2	36.1	34.4	0.8	-2.1	2.9 (B)
Earnings loss due to punishment	-3.3	-5.3	-0.7	-3.3	-2.6	-2.0	-0.6 (C)
Of which: Earnings loss due to punishment of cooperators displayed as free riders (factor 1)	0.0	-0.7	0.0	-1.4	0.0	0.7	-0.7 (D)

Panel B

Decomposition of drop in earnings under IS relative to FS when noise is introduced	Absolute	Relative (percent)
Factor (1) (D)	-0.7	-29.8
Factor (2) (B + C - D)	3.0	129.8
Total difference in effect of noise (A)	2.3	100.0

Note: Observations from all groups using IS/FS, respectively, are used. The labels A – D in Panel B refer to the same labels in the last column of Panel A. In the "Earnings before punishment" row, the fixed fee (c = 3 points) in FS is subtracted.

The first row of the last column ("difference in differences") in Panel A shows that the introduction of noise caused a 2.3 points larger drop in earnings in IS than in FS, in terms of experimental points per person per period. The second row of the same column shows that this difference is more than fully accounted for by differences in earnings *before* punishment, i.e. in earnings calculated before the cost of punishment has been subtracted. The introduction of noise led to more punishment in both IS and FS but this increase was stronger in FS than in IS, as shown in the third row. In particular, the average income loss due to punishment of cooperators displayed as free riders, i.e. factor (1), is 0.7 points per person per period in IS with noise and 1.4 points per person per period in FS with noise. The amount is higher in FS than in IS because all cooperators, who were displayed as free riders, were punished in FS, whereas only a fraction (72 percent) received punishment in IS. Moreover, punishment was more severe in FS (16 points per person per incidence) than in IS (8.8 points). This analysis only accounts for the direct, first-order effect of punishment on earnings. In theory, there could be additional effects if punishment of a cooperator displayed as a free rider leads him or her to free ride in the following periods. However, the results presented in Table D2 suggest that such effects are weak.[29]

[29] I.e. the positive effect of "punishment points received*type I error$_{t-1}$" in model 4 and the insignificant effect of "type I error$_{t-1}$" in model 8.