Kameik, Kenju; Putterman, Louis

Working Paper

In broad daylight: Full information and higher-order punishment opportunities promote cooperation

Working Paper, No. 2012-3

Provided in Cooperation with:
Department of Economics, Brown University

Suggested Citation: Kameik, Kenju; Putterman, Louis (2012) : In broad daylight: Full information and higher-order punishment opportunities promote cooperation, Working Paper, No. 2012-3, Brown University, Department of Economics, Providence, RI

This Version is available at:
https://hdl.handle.net/10419/62671

In Broad Daylight: Full Information and Higher-order Punishment Opportunities Promote Cooperation

Kenju Kamei[1], Louis Putterman[2,*]

[1] Department of Economics, Bowling Green State University, Bowling Green, OH 43403, USA. Email: kenju.kamei@gmail.com.

[2] Department of Economics, Brown University, 64 Waterman Street, Providence, RI 02912, USA. Email: Louis_Putterman@brown.edu.

* Corresponding author: Louis_Putterman@brown.edu. Tel: +1 (401) 863-3837. Fax: +1 (401) 863-1970.

Abstract:

The expectation that non-cooperators will be punished can help to sustain cooperation, but there are competing claims about whether opportunities to engage in higher-order punishment (punishing punishment or failure to punish) help or undermine cooperation in social dilemmas. In a set of experimental treatments, we find that availability of higher-order punishment increases cooperation and efficiency when subjects have full information on the pattern of punishing, including its past history, and opportunities to punish are unrestricted. Availability of higher-order punishment reduces cooperation and efficiency if it is restricted to counter-punishing alone, if past history is unavailable, and if there is a dedicated counter-punishment stage.

Keywords: collective action, social dilemma, voluntary contribution, public goods, punishment, counter-punishment, higher-order punishment.

JEL classification codes: C9, H41, D0

Research Highlight:

- We conduct voluntary contribution experiments with opportunities to punish both conditional on others' contributions and conditional on others' punishments.
- We find that higher-order punishing opportunities increase cooperation and earnings when subjects learn of all punishments, are shown histories of past decisions, and can engage in higher-order punishment of any group member.
- We find that higher-order punishing opportunities reduce cooperation and earnings when subjects learn only who punished them (ego-centric information), see no history of past decisions, and higher-order punishment is limited to counter-punishing.
- Concern that knowing who punished whom and having opportunities to retaliate will undermine voluntary collective action is found to be unwarranted under conditions of symmetric information and punishment opportunities.

1. Introduction

In the growing body of theoretical, field, and experimental research on cooperation in social dilemmas, the role of punishment has received considerable attention. Many subjects in experiments are seen to engage in costly punishment even in the absence of strategic motives for doing so (Fehr and Gächter, 2002; Falk *et al.*, 2005). In subject pools drawn from societies with well-functioning institutions, most punishment is directed at non-cooperators, and the availability of punishment leads to higher cooperation levels (Herrmann *et al.*, 2008). But several questions remain unsettled, including what motivates punishment, and whether the benefits of offering punishment opportunities can survive opportunities to counter-punish.

One proposed explanation of the propensity to punish is that a preference for punishing non-cooperators (that is, for engaging in first-order punishment) could have been evolutionarily selected for thanks to second-order punishment of those who failed to (first-order) punish. If further enforcement steps were universal up to some n^{th} order of punishment, the need to punish at that stage might be invoked so rarely that the payoff disadvantage of an n^{th} order punisher would be swamped by the advantages shared by all members of groups in which punishing types predominate (Henrich and Boyd, 2001; Henrich, 2004).[1] Axelrod (1986) discusses "a norm that one must punish those who do not punish a defection," labeling it a "meta-norm." These discussions suggest that higher-order punishment is helpful, and perhaps even necessary, for fostering cooperation.

Recently, some economists have viewed higher-order punishments as problematic rather than helpful, however. Their concern harks back to John Locke's (2005 [1739]) argument that sanctioning should be the province of government rather than of individual citizens because individuals are reluctant to punish due to the danger of counter-punishment. Locke asserted that "resistance many times makes the punishment

[1] On the reintroduction of group selection into the literature on evolutionary theory, see the discussion in Henrich (2004) and sources cited there including Sober and Wilson (1998).

dangerous, and frequently destructive, to those who attempt it," and that people therefore willingly cede their rights to punish individually to the state, which punishes on their behalf. The potential of counter-punishment to deter and thus to undermine the efficacy of punishment, while adding to its cost, has recently been demonstrated in laboratory experiments by Denant-Boemont *et al*. (2007), Engel *et al.* (2011), Nicklisch and Wolff (2011), Nikiforakis (2008), and Nikiforakis and Engelmann (2011). Nikiforakis (2008) suggests that the problem may be a fundamental one, sub-titling his paper "Can we really govern ourselves?"—a rejoinder to Ostrom *et al.*'s (1992) sub-title "Self governance is possible." Counter-punishment is also related to perverse or anti-social punishment—i.e., punishment of high contributors or cooperators—as indicated by the fact that most such punishments appear to be attempts at "blind revenge" (Cinyabuguma *et al.*, 2004; Herrmann *et al.*, 2008).[2]

In addition to the possibilities that higher-order punishment opportunities will be used to punish those who fail to do their part in punishing norm-violators (as suggested by Henrich and Boyd) or that it will be used to retaliate against the punisher (as emphasized by Nikiforakis and others), pro-social actors might use higher-order punishment opportunities to punish those who punish cooperators rather than non-cooperators at the initial opportunity to punish. Such pro-social higher-order punishment is documented in experiments by Cinyabuguma *et al.* (2006) and by Denant-Boemont *et al.* (2007), the latter grouping it along with punishment of non-punishers in what they call "sanction enforcement." In what follows, we'll refer to the punishment of (first-order)

[2] Bochet *et al*. (2006) coined the term "perverse punishment" to refer to cases in which a subject who contributes above his group's average in a period, and especially one contributing the maximum observed amount in the group, is punished. Cinyabuguma *et al*. (2004) confirm the conjecture that when highest contributors are punished it leads them to reduce their contributions, demonstrating that such punishment is "perverse" in the sense that it is efficiency-reducing rather than efficiency-enhancing. Herrmann *et al*. parse their data somewhat differently, labeling as "anti-social punishment" instances in which a group member punishes someone who contributes more than herself. Bochet *et al*. and Cinyabuguma *et al*. prefer to define "perverse punishment" with reference to the recipient's contribution only, rather than the comparison of the recipient's with the punisher's contribution, because in most of the experiments in question the recipient does not learn who the punisher was, so the incentive effect of the punishment can be affected only by the recipient's contribution. Both sets of researchers agree that in practice the large majority of cases satisfying the definition of "perverse punishment" likewise would be classified as "anti-social punishment."

non-punishers and the punishment of (first-order) *perverse punishers* as PEO (punishment enforcement for omission) and PEC (punishment enforcement for commission), respectively.

To better understand whether opportunities to engage in higher-order punishment are helpful or harmful to cooperation, we conducted a series of experiments in which we varied the number of opportunities to punish, the information available at each punishment stage, and who subjects are permitted to punish when an additional punishment stage is included. Like the work cited above, our starting point is a multi-player, finitely repeated linear voluntary contribution mechanism (VCM, also known as public goods game) modified so that each period includes a post-contribution stage in which group members learn one another's contributions to the public good (group account) and have the chance to punish one another at some cost. In the standard design (e.g., Fehr and Gächter, 2000), group members are not informed of who punished them, and identifiers are scrambled each period, to avoid vendettas. We conduct a reference treatment having this standard design, and we conduct additional treatments to study the effect of opportunities to engage in higher-order punishment.

Our additional treatments differ in three dimensions. First, in some treatments, information about punishments given and opportunities to engage in higher-order punishment are restricted to knowledge of who punished oneself and opportunities to counter-punish. We refer to these treatments as having an "ego-centric" structure of information and of higher-order punishment opportunities. Other treatments are not so restricted but rather include full information about all punishments in the group and opportunities to punish any group member one wishes to. Our second dimension of treatment variation concerns whether there is or is not a distinct stage each period dedicated to higher-order punishment. Whereas most of the new treatments we study include a distinct third stage in each period, we also study two treatments without such a stage. These allow higher-order punishment of period t punishing behaviors in period t+1, but of necessity such punishing must be simultaneous with first-order punishment of

period t+1 contribution decisions. Finally, our treatments vary with respect to whether information regarding past punishments and contributions is displayed in later periods, as opposed to each period's information being self-contained.

To foreshadow results, in all but one of our new treatments, providing details about who punished whom how much, along with opportunities to engage in higher-order punishment, prove unharmful to achieved levels of cooperation and efficiency, and in at least one treatment this information and the associated higher-order punishment opportunities are distinctly helpful. The treatment yielding clearly higher contributions and earnings than the standard reference treatment is one with symmetric information, generalized higher-order punishment opportunities, a dedicated higher-order punishment stage, and carry-over of information on past behaviors. The sole treatment in which information and additional punishment opportunities prove harmful is one in which information and higher-order punishment opportunities are ego-centric, there is no carry-over of history, and there is a dedicated counter-punishment stage.

While more generalized information and opportunities to engage in higher-order punishment lead to more rather than less cooperation, the data provide little evidence that this is due to pro-social punishment being rendered more common because abstaining from it is punished (PEO), as in the Henrich-Boyd scenario. Nor is there much direct evidence of PEC (punishment enforcement for commission) of the kind discussed by Denant-Boemont *et al*. Rather, we see a more pro-social pattern of punishment, including significantly less counter-punishing of first-order punishment aimed at low contributors than of that aimed at high ones, in treatments in which punishment information and higher-order punishment opportunities are symmetric than in those in which they are ego-centric. It seems that common knowledge of the pattern of contributing and punishing—the "sunshine" of more complete information—may be as important as are higher-order disciplinary opportunities themselves for fostering cooperation.

The remainder of our paper proceeds as follows. Section 2 provides additional background and details of our experimental design. Section 3 discusses our experimental results. Section 4 summarizes our conclusions.

2. Background and experimental design

2.1 Literature and design considerations

In first-generation laboratory experiments in which punishment opportunities are added to a linear voluntary contribution mechanism (Fehr and Gächter, 2000; Fehr and Gächter, 2002; Masclet *et al.*, 2003; Sefton *et al.*, 2007; Carpenter, 2007; Bochet *et al.*, 2006), subjects are provided with an endowment of experimental currency in each period and simultaneously make first-stage decisions on what, if anything, to contribute to a group account. Each period includes a second stage in which each subject is shown the first-stage contributions of each of the others and decides how much (if any) costly punishment to give. At the end of the period, each subject learns how much punishment she received but not which group members in particular punished her how much. Subject *i*'s earnings in period *t* are given by

$$\left\{E - C_{it} + r \cdot \sum_{j=1}^{n} C_{jt}\right\} - \beta \cdot \sum_{j=1, j\neq i}^{n} p_{ji}^{t} - \sum_{j=1, j\neq i}^{n} p_{ij}^{t}, \quad (1)$$

where E is the per-period endowment common to all subjects, C_{it} is subject i's allocation to the public good in period t, n is the number of group members, r is the marginal per-capita return (MPCR) per unit allocated to the public account, and p_{ji}^{t} is the number of units of punishment subject j gives to subject i in period t. Here, the term in the curly bracket is subject i's earnings from the allocation stage, the second term, $\beta \cdot \sum_{j=1, j\neq i}^{n} p_{ji}^{t}$, is her loss due to receiving punishment from other members (with β being the loss to the targeted individual per point of punishment given), and the third term, $\sum_{j=1, j\neq i}^{n} p_{ij}^{t}$, is her expense to give punishment to others. Setting $1/n < r < 1$ assures that the underlying game is a social dilemma since the social optimum entails $C_{it} = E$ for all i and all t

whereas maximization of own payoff taking others' contributions as given entails $C_{it} = 0$ for all *i* and all *t*. In one-shot play or in the last period (which is indicated by "*T*") of finitely-repeated play, private payoff-maximizing behavior entails $p_{ji}^{T} = 0$ for all *j* and *i*, so threats to punish in earlier periods are not credible if there is common knowledge that all group members are rational maximizers of own payoff.

In dozens of past finitely repeated VCM experiments *without* punishment opportunities, the prediction that $C_{it} = 0$ fails, with contribution averaging between 40 and 60% of endowment in the initial period. Contributions then decline more or less monotonically with repetition (Zelmer, 2003). When costly punishment is available, enough subjects pay for punishment and it is sufficiently well targeted at lower contributors in typical experiments that contributions decline more slowly or even rise with repetition, the rising trend being the more likely the higher is β (Nikiforakis and Normann, 2008) and the more norm-following is the institutional milieu of the subject pool (Herrmann *et al*., 2008). Repeated play is typically used in these experiments in order to compare change over time in punishment conditions with that in the VCM without punishment. The number of repetitions is announced in advance so that there is a straightforward prediction of no contributions and no punishment under classical assumptions. In actuality, contributions often fall in the last period, presumably because some subjects guess that punishment is unlikely then, but in fact a given deviation of contribution below the group average tends to be punished at least as much in that period, indicating that punishment is not, after all, primarily strategically motivated (Falk *et al.*, 2005).

In the experiments referred to, subjects are not informed who punished them by what amount and identifiers are switched from period to period to discourage vendettas of counter-punishment. Subjects also lack information about others' punishing practices, which along with the identification changes means that PEO and PEC are ruled out. Some subjects do attempt to counter-punish—e.g., a low contributor punished in period *t* may punish a high contributor in period *t* + 1 in the belief that that group member is

likely to be a perennial high contributor and that it is high contributors who punish low ones. The absence of proof causes such instances to be labeled "blind revenge" by Ostrom *et al* (1992). Cinyabuguma *et al*. (2006) and Herrmann *et al*. (2008) judge this to be the likely main cause of observed perverse and anti-social punishment. One can conjecture that openly providing the information on which counter-punishment can be based would lead to more such revenge, which would directly lower efficiency, since both punisher and punishee lose resources. It might also reduce contributions, since the anticipation of counter-punishment could deter first-order punishing (as in the Locke quotation). The results obtained in some of the treatments in Denant-Boemont *et al*. (2007), Engel *et al.* (2011), Nicklisch and Wolff (2011), Nikiforakis (2008), and Nikiforakis and Engelmann (2011) support this conjecture.

But the experiments just mentioned have their own restrictions. In some treatments such as the counter-punishment treatment in Nikiforakis (2008) and the "revenge only" treatment in Denant-Boemont *et al*. (2007), subjects learn only of punishments directed at them and in the additional stage decide only how much to punish back, which runs the risk of engendering an "experimenter demand effect" (Zizzo, 2010).[3] Other papers have studied treatments permitting higher-order punishment that is *not* restricted to counter-punishing. Cinyabuguma *et al*. (2006) periodically reported others' punishing histories to each player in an unidentified fashion showing only how much punishment each had given to those contributing above the average in their group, those contributing below that average, and those contributing exactly the average. Subjects could then engage in costly punishment on the basis of that categorized first-order punishment information. They found substantial willingness to pay for second-order punishment, with first-order punishers of above-average contributors receiving about three times as much second-order punishment as first-order punishers of below-average contributors, and with first-order non-punishers (the omitted category) being punished the least. Although first-order punishment of above-average contributors

[3] Engel *et al*. (2011) also replicate the Nikiforakis (2008) counter-punishment treatment, while providing subjects in some treatments information on past play to study the manipulation of prior expectations.

accordingly declined, punishment of below-average contributors also declined slightly. Overall, contributions and earnings were slightly but not statistically significantly higher than in a comparison treatment without higher-order punishment.

While Denant-Boemont *et al*. (2007)'s "revenge only" treatment replicates the counter-punishment treatment in Nikiforakis (2008), they also conduct treatments in which subjects receive more complete information on the current-period punishments between all pairs of group members and have an additional opportunity to punish any group member at the same cost ratio as with first-order punishment. In the retaliation-only treatment, their findings are largely in line with those of Nikiforakis: the contribution level fails to rise with repetition, and it is significantly lower than in the basic punishment treatment (with only one punishment stage). In their full information treatment with a complete range of higher-order punishment opportunities, average contributions fall somewhere between those in the basic punishment treatment and those in the counter-punishment treatment, with a mildly rising trend, and a difference from those in the basic treatment significant at the 10% level. Denant-Boemont *et al*. also study a treatment with full information in which another four punishment stages are available each period, and Nikiforakis and Engelmann (2011) study a similar treatment in which the number of higher-order punishment stages is determined endogenously in each period.[4]

We design our new treatments with particular attention to the concern that the apparent power of peer-to-peer punishment to stabilize cooperation in first-generation cooperation and punishment experiments like Fehr and Gächter's is misleading because such experiments artificially shield subjects from the consequences of punishing one another. That shielding would come from two main treatment features: (1) the fact that

[4] In Nikiforakis and Engelmann, subjects are given the full range of punishment decisions in each of their punishment stages and can engage in higher-order punishment of any kind, but, unlike Denant-Boemont *et al.*'s six stage treatment, the number of punishment stages is determined endogenously. If at least one subject in a group assigns positive punishment points to another member and at least two group members still have positive earnings for the period, the group moves on to another stage in which additional punishments can be assigned by those able to pay for them. Nikiforakis *et al.* (forthcoming) study similar treatments in a setting with heterogeneous returns from the public good.

subjects never learn exactly who punished them and how much, and (2) the fact that subjects are prevented from punishing back—other than "blindly"—by the scrambling of identifiers and absence of historical reminders. We also address the concern that the first generation experiments rule out PEO and PEC.

More than one new treatment is required because there is room for debate about exactly how the shielding and limitations in question should be removed so as to allow for a more realistic and complete view of informal sanctions in the real world. Our treatment decisions are made clearer by considering three sets of issues, as follows.

History and identifiability. In ongoing real world interactions, information might be recalled and might then influence future punishments any time after an individual becomes aware of a punishment event. Effects are likely to decline as memories recede, so presence or absence of reminders may be important. Informed higher-order punishments are impossible beyond the current period in experiments in which identities are scrambled, unless the relevant information is specially presented (meaning that past punishment actions are displayed with other information about a subject despite absence of a fixed position or numerical or letter identifier of that subject on the screen). We vary history and identifiability along the spectrum from conditions resembling Fehr and Gächter (2000) and Nikiforakis (2008) (i.e., neither identifiability nor display of history beyond the current period) to ones in which some historical information is presented in future periods or subjects keep fixed identifiers across periods, and finally to ones with both fixed identities and repeated display of information to aid memory.

Information and punishment restrictions. As mentioned, some recent experiments restrict information on punishment and opportunities for higher-order punishment in a manner we describe as "ego-centric," and this rules out both PEO and PEC. A justification for the ego-centric approach may be that, especially in larger groups, one may be able to readily observe punishments to and from oneself only. But the conditions of observability are situation-specific, making it reasonable to consider the ego-centric restrictions (only j observes p_{ij}) as one end of a continuum of possibilities. We envision

equal observability of punishment interactions between any two group members (*i* and *j*, *j* and *k*, etc.) as lying at the other end of that continuum, with many real-world situations lying in between.[5] In our experiment, we explore the two extremes only, keeping in mind that neither is likely to perfectly represent most realities.

Stages and punishment opportunities. One way to study higher-order punishment in the lab is to give subjects opportunities to engage in it at designated decision stages. But thinking of higher orders of punishment as occurring at discrete points in time, and in sequence, is a convenience that might easily be carried to unrealistic extremes. Whereas in ongoing interactions it is entirely plausible that punishment chains reaching up to very high orders may take place, it strikes us as unreasonable to model this in the lab as a sequence of distinct punishment stages, as happens in treatments of Denant-Boemont *et al*. (2007) and Nikiforakis and Engelmann (2011) in which as many as 6 or 7 acts of counter-punishment, counter-counter-punishment, etc., take place in a single period. One might instead think of individuals as allocating resources to two activities—contributing (or not) to the public good, and punishing (or not punishing) others, and recognize that acts of punishing should not necessarily fill more temporal slots than are accorded to acts of contributing, and that giving subjects more opportunities for punishing than for contributing might induce an experimenter demand effect.[6] Having exactly one extra stage—the third stage of the period—available for higher-order punishment—as in the

[5] The observability of peer-to-peer punishments in the real world depends, among other things, on the physical details of interactions and the available channels of communication. In a small group working together in a workshop or field of modest size, interactions among pairs of others may be almost as well observed as interactions that include oneself, but this is hardly the case if we're considering by-stander monitoring of acts such as littering in a large society, where only those in the immediate vicinity are in a position to show disapproval. Note that perfect observability of punishment directed at oneself also cannot be assumed in all situations. For example, many real-world punishments take the form of being gossiped about "behind one's back," with consequent loss in esteem, but the subject of such gossip may not be able to tell exactly who in the group has bad-mouthed him or her to what degree.

[6] The concern is that the subject is being paid by the experimenter to participate in an experiment, and that the subject has no other task with which to occupy her attention than that of making whatever decision the experiment calls for. The more times an experiment asks subjects to decide how many points to give to punishment, the more total punishment subjects might give, despite the fact that zero punishment is always one of their options. The concern about experimenter demand is arguably mitigated by having the number of punishment stages be determined endogenously in Nikiforakis and Engelmann (2011) and in Nikiforakis *et al*. (forthcoming), but the possibility of imbalance between attention to punishing decisions and attention to contributing decisions may still be an issue.

punishment ("revenge only") treatments of Nikiforakis (2008) and Denant-Boemont *et al*. (2007) as well as the latter's "no revenge" and "full information" treatments—can be defended as a reasonable compromise. Letting subjects be reminded of past history while still having only the same number of discrete opportunities to punish as to contribute to the public good—hence two stages per period—is an alternative way to allow higher-order punishment, and perhaps one with less cueing of subjects towards it. We experiment with each of the two approaches.

2.2 The experiment

Our experiment includes a Reference treatment of the standard first-generation cooperation and punishment variety, with a single punishment opportunity and scrambling of identifiers. Along with it, we conduct six treatments providing higher-order punishment opportunities, as summarized in Table 1. In each session, sixteen or twenty undergraduate participants are randomly and anonymously assigned to groups of four who interact without change of partners for a total of fifteen periods. Subjects are clearly told that the experiment will be over in fifteen periods. To simplify instructions and interpretation, we use a fixed ratio of punishee loss to punisher cost (parameter β of Eq. (1) above) rather than the rising marginal cost to punisher and fixed percentage loss to punishee used by Nikiforakis and Denant-Boemont *et al*.[7] The punisher pays one point to reduce the earnings of the targeted individual by three points ($\beta = 3$).[8] To avoid the possibility that subjects have to pay the experimenter for losses, we constrain earnings net

[7] Those authors adopt the punishment cost structure of Fehr and Gächter (2000), in which increasingly expensive punishment points each deprive the targeted individual of 10% of her pre-punishment earnings for the period. This has the perhaps undesirable consequence that a punishment point takes more from a low than from a high contributor, which could bias targeting of punishment towards "free riders" if punishers want maximum "bang for their buck;" it also makes subjects' calculations more difficult. The fixed punisher-to-punishee cost ratio used by us has become common in the literature, for example Fehr and Gächter (2002), Page, Putterman and Unel (2005), Bochet, Page and Putterman (2006), and Nikiforakis and Normann (2008).

[8] The 1:3 ratio is used in other experiments including Fehr and Gächter (2002). Nikiforakis and Norman (2008) compare the efficacy of 1:1, 1:2, 1:3 and 1:4 ratios, and find that a ratio of at least 1:3 is required to prevent contributions from declining with repetition.

of punishment incurred to be non-negative, but to assure that punishing is always costly and hence a non-payoff-maximizing action under traditional assumptions (i.e., common knowledge of rational maximization of own payoffs), subjects always incur the cost of any punishing they themselves chose to impose. Net losses, in practice rare and limited to a few periods, were covered out of earnings from other periods.[9] Earnings in a given period are accordingly:

$$max\left\{\left\{20 - C_{it} + 0.4 \cdot \sum_{j=1}^{n} C_{jt}\right\} - 3 \cdot \left(\sum_{j=1, j\neq i}^{n} p_{ji}^{t} + \sum_{S_j^t} pp_{ji}^{t}\right), 0\right\} - \sum_{j=1, j\neq i}^{n} p_{ij}^{t} - \sum_{S_i^t} pp_{ij}^{t}, \quad (2)$$

where pp_{ji}^{t} is subject *j*'s punishment of subject *i* due to *i*'s punishing behavior, and S_j^t is the set of subjects about whom subject *j* is provided with information on which to base higher-order punishment.

In three treatments, dubbed E2, E3n and E3h, where E indicates ego-centric information, each subject learns only who punished himself or herself by how many points, while in three counterpart treatments, dubbed F2, F3n and F3h, where F indicates full information, subjects learn the amounts of all bilateral punishments within the group. Four treatments—E3n, E3h, F3n and F3h—add a second opportunity to punish each period, hence they have three stages (contribution, first punishment stage, second punishment stage). Two treatments—E2 and F2—do not add an extra stage. In E2, each subject knows who punished him or her by how much in the previous period (say, period *t*) when deciding on punishment after the contribution stage of the next period (say, *t*+1). The subject can condition punishments in period *t*+1 on both current contribution and past punishment by the person targeted. F2 is set up similarly, but in F2 subjects have information about all group members' punishments of one another, so the punishment

[9] The constraint that first-stage earnings minus punishment received cannot fall below zero was binding in 23 out of 4,080 periods of individual subject play in the seven treatments studied. Earnings after deduction of costs to punish others were negative in 22 periods out of the same number of periods of individual play.

stage can take into account not only others' contributions and punishment of oneself but also punishment of others, which makes PEO and PEC possible.

Following Fehr and Gächter (2000), Nikiforakis (2008), and other papers, subject identifiers are scrambled each period in the three E treatments. Counter-punishment is made possible in E2 by a special display of information about each group member's past punishment of oneself. In E3n, no individually-linked information survives beyond the current period, exactly as in Nikiforakis (2008) and Denant-Boemont *et al*. (2007). E3h has three stages per period, like E3n, but provides information on past actions, like E2. (The names of the three-stage treatments are distinguished by an h for history or an n for no history.) If differences between E2 and E3n are mainly attributable to the displaying of past information in E2, behaviors in E2 should resemble those in E3h; but if differences between E2 and E3n are mainly due to the difference in the timing of decisions (the number of stages) and so perhaps to experimenter demand effect, behaviors in E2 and E3h may differ. Differences in behaviors between E3n and E3h should be attributable only to the presence of historical information in the latter.

Because the spirit of the F treatments is one of full information, subject identifiers and screen positions are not scrambled from period to period in them. The difference between F3n and F3h, therefore, is not properly speaking that more information about past behaviors is made known to subjects in F3h than in F3n, but that subjects receive ongoing reminders of past punishments and contributions of the others in their group in F3h, but would need excellent recall to remember with comparable detail in F3n. As mentioned already, F2 is a two-stage treatment in which higher-order punishment can take place simultaneously with punishment conditioned on contribution decisions, and unlike E2, that higher-order punishment is not restricted to counter-punishment but can include PEO and PEC.

In treatments having three stages per period, we can think of the p_{ij}^t terms of Eq. (2) as indicating punishments in stage 2 (the first punishment stage) and the pp_{ij}^t terms as indicating those in stage 3 (the second punishment stage). The payoff function in the two

stage treatments can also be rendered by Eq. (2), but since each subject *i* submits only one number indicating the punishment points she gives to each *j*, we cannot perfectly distinguish, observationally, between *p* and *pp* in these treatments. We will nevertheless tease out plausible inferences about first versus second-order punishment in the two-stage treatments, as will be seen in some of the regression analysis of Section 3.

In both the two and the three-stage treatments with history information, we chose to display past behaviors not only from the most recent period, but also a cumulative average up to the previous period. We decided to provide more information than might minimally be needed for a single round of higher-order punishment in the treatments with history display because we wanted to study whether such information might prove helpful in its own right, for instance facilitating the emergence of norms of cooperation via the fuller and more salient display of information.

Table 2 provides details about the information available to subjects at the various stages of a period, by treatment. Experiment instructions are included in the Online Appendix.

3. Results

14 experiment sessions, 2 for each treatment, were conducted in a computer lab at Brown University between October, 2011 and January, 2012. Participants were recruited from the general undergraduate population, representing majors in the humanities, social sciences, and sciences, with 18% being economics concentrators (slightly higher than their share in the general student population) and 51.5% female (almost perfectly representative).[10] The large majority had no previous experience of a public goods experiment, and each participated in one session only. Sessions typically took 75 to 90

[10] Students responding to flyers register as potential participants in the BUSSEL (Brown University Social Science Experimental Laboratory) data base, modified from CASSEL (California Social Science Experimental Laboratory), and respond to email messages inviting their participation at specific dates and times. The messages indicate that participants are guaranteed a $5 show-up fee and will earn an unspecified additional amount "usually averaging between $15 and $25."

minutes from signing of consent forms to reading aloud and (simultaneously) on paper of instructions, answering of comprehension questions, engaging in the fifteen decision periods, and privately receiving cash payment, which averaged $20.13 (1 experimental point = $0.05) plus a $5 show-up fee.

Fig. 2 displays the trends of average contributions period by period for each treatment. The Reference treatment displays the typical pattern of contributions in first-generation contribution and punishment treatments. The average contribution begins a little below 60% of endowment, and then trends upwards towards 75% of endowment before a last-period decline.

The first important thing to notice about Fig. 2 is that although every treatment other than Reference offers subjects the opportunity to engage at least in counter-punishment and possibly in other kinds of higher-order punishment, all but one of the six treatments shows no sign of contributions being lower than in Reference. Average contribution is higher than in Reference (although not necessarily significantly so) in every period for three treatments, and in the majority of periods for another two treatments. However, using group-level observations of average contribution for periods 1 – 15 as a whole, Mann-Whitney tests find that the distribution of contributions is statistically significantly different from Reference only for the treatment having the highest average contribution curve, F3h ($p = .014 < 0.05$, 2-tailed test).

In one treatment, contributions are clearly *lower* than in Reference. That treatment is E3n, the ego-centric information treatment modeled on those of past counter-punishment experiments including Nikiforakis (2008) and Denant-Boemont *et al.*'s "revenge only" treatment. Fig. 2 shows E3n contributions having an initial contribution uptick followed by persistent decline from periods 5 to 15. Average contribution for the 15 periods as a whole is statistically significantly lower in E3n than in all other

treatments except Reference.[11] While the difference between contributions in E3n and Reference for the full 15 periods is significant at the 10% level only in a one-tailed test (p = .060), group average contributions in E3n differ significantly from those in Reference when periods 8 – 15 alone are considered (p = .023 in a two-tailed test). The trends of average earnings as well as significances of differences in average earnings between treatments are similar to those for average contributions. Details are provided in the Online Appendix.

To explain the differences in contribution patterns, we looked at differences in the use of punishment opportunities, including differences in the extent and targeting of first-order punishment, and frequency of counter-punishment, punishment of non-punishers, and punishment enforcement. Stage 2 behaviors are ostensibly similar in all treatments: 71.7% of subjects punished at least once, 70.8% were punished at least once, and the average subject punished at least one other subject in 17% of periods. Out of the three opportunities available to (first-order) punish other group members each period, about 3.9% were used to assign a positive amount of punishment in F3n, 5.2% in E3h and F3h, and 7.9% in E3n. The corresponding shares are 9.7% and 8.4% in E2 and F2, respectively. 83.1% of second stage punishment was targeted at low contributors, with small differences in targeting across treatments.

Estimates of regression equations following a specification in Fehr and Gächter (2000) find that, as with their data, the further below his group's average contribution in a given period was a subject's own contribution, the more punishment he received, significant at the 1% level. Raising his contribution further above the group average, on the other hand, left punishment unaffected in most treatments.[12] This indicates that while

[11] In two-tailed Mann-Whitney tests with group level observations for periods 1 – 15, contributions in E3n differ significantly from those in E2, F2, and F3h at the 1% level, from those of F3n at the 5% level, and from those of E3h at the 10% level.

[12] See Online Appendix, Table B.3. A separate regression is estimated for each treatment. The coefficient on the negative deviation term falls short of the 1% significance level but is significant at the 5% level in the regression for the E3n treatment. In two of the treatments, F3n and E3h, there is a significant although small positive coefficient on the Positive Deviation term, suggesting that contributing too much *above* the

motives to punish free riding may be mixed with motives to counter-punish, especially in treatments E2 and F2, the predominant motive of punishers seems the same as in other experiments. Coefficient values range from 0.16 to 0.44, meaning average punishment received for contributing one less point was less than the 0.6 required to render contributing privately profitable.[13]

We begin looking for evidence of higher-order punishment in the two-stage treatments, where it is in principle more difficult to find due to the combining of first- and second-order punishment in a single punishment stage. Table 3 shows estimates of regressions resembling those just described, where the dependent variable is punishment received by subject *j* in period *t*, but explanatory variables are added for amount of 2nd stage (first-order) punishment *j* gave to below-average contributors in the previous period and the corresponding amount *j* gave to above or equal to average contributors in that period. (These regressions exclude period 1 observations, since in that period no previous period information was available.) The coefficients on the negative deviation of period *t* contribution term remains significant and little changed, but one of the new terms added to pick up possible punishment for punishing last period obtains a significant coefficient in each regression. This confirms the possibility of detecting counter-punishment despite its temporal mingling with first-order punishment, i.e., that conditioned on current contribution.

Importantly, we encounter our first important sign of asymmetry between behaviors in the E and F treatments, here: while it is the coefficient on last period punishment to low contributors that obtains a significant positive coefficient in the E2 regression, the one on last period punishment given to above- or equal-to-average contributors has the significant positive coefficient in the F2 regression. This means that in the ego-centric treatment, E2, there is significant counter-punishing of those who pro-socially punished in the previous period, whereas in the full information treatment, F2,

---

average attracted perverse or anti-social punishment. Önes and Putterman (2007, Table 2) find similar results for some treatments.

[13] Point estimates are 0.37 and 0.33 in treatments E2 and F2, giving little support for the idea that punishment of free riders is any less important in them.

there is significant counter-punishing of those who perversely or anti-socially punished high contributors. Put differently, second-order punishment is relatively "anti-social" in E2 and relatively "pro-social" in F2.

Turning to the three stage treatments and focusing still on counter-punishment (punishing back by the recipient of first-order punishment), Fig. 3 shows the fraction of second-stage punishment events ($p_{ij}^t > 0$) that are followed by third-stage counter-punishment from the recipient ($pp_{ji}^t > 0$). The proportion of cases that apparently attract counter-punishment is substantial, in the 20 to 50% range in most cases. Importantly, *perverse* first-order (second-stage) punishments are met by retaliation in a much higher proportion of cases than are pro-social ones in the F but not the E treatments, much as Table 3's regressions suggest for the two-stage treatments. Both sets of results suggest a greater tendency to abide by pro-social norms when fuller information is available. Retaliation against pro-social second-stage punishers is especially low in F3h—the treatment that attains highest efficiency. There, only 7% of pro-social punishment events are followed by counter-punishment.

The tendency for the ratio of "pro-social" to "anti-social" higher-order punishment to be greater in F than in E treatments is also found in regressions that investigate the incidence of all higher-order punishment, including but (in F treatments) not limited to counter-punishment by the person initially targeted. In Table 4, we treat third-stage punishment received by subject *j* as being a function of *j*'s second-stage punishment of low and of high contributors in the same period. All point estimates except those for E3n indicate that there was more 3rd stage punishment of perverse than of pro-social 2nd-stage (first-order) punishers. But in the F treatments, only the coefficients on perverse second-stage punishment given are statistically significant, indicating that there was a significant tendency to use third-stage punishment to support pro-social norms there. In E3n, the counter-punishment treatment that shows worst performance overall, neither coefficient is significant, but the point estimates imply that there was if anything more 3rd-stage punishment of an "anti-social" kind—i.e.,

punishment given to those who engaged in pro-social punishment at the second stage—than 3rd-stage punishment of a "pro-social" kind. Finally, while the regression for E3h shows significant counter-punishment of pro-social 2nd-stage punishers, it also shows far more counter-punishment per point of anti-social punishment, which may help to explain why cooperation was much higher in E3h than in E3n. Conceivably the ongoing display of past behaviors in E3h helped cooperatively-oriented subjects to understand the dilemma more clearly and so to act more forcefully to resist intimidation by those whom they (first-order) punished, even though the information given had the ego-centric limitation common to the E treatments.

We looked in our three-stage treatments, where the evidence should be clearest, for signs of both PEO (punishment enforcement for omission) and PEC (punishment enforcement for commission). In F3n and F3h, we found cases in which a second-stage non-punisher received punishment in the third stage, consistent with PEO. But almost all of these cases can be explained as delayed first-order punishment.[14] To check for PEC, we identified all cases in which an equal-to-or-above-average contributor was (perversely) punished in the second stage and there was a third group member, not the punished subject, who had an opportunity to punish the 2nd-stage punisher responsible. Limiting this search to potential third-party punishers who were themselves above-average contributors, we found only five and eleven such opportunities in F3n and F3h, respectively. Out of these potential cases, we found actual 3rd-stage punishment in none of the F3h and in only one of the F3n cases. So PEC also appears to be rare, in our data.

[14] The targeted individuals were low contributors, and by delaying punishment to stage 3, the punisher eluded counter-punishment, especially in F3n with its lack of history display. One form of indirect evidence about what was in fact motivating some subjects to punish individuals who failed to punish in the second stage is to see how the punishment recipients themselves responded to being punished. Regressions shown in Appendix Table B.4 find evidence that such recipients of third-stage punishment responded by raising their contributions, not by engaging in the second- stage punishing on which they had "shirked." Thus, the punished subjects themselves appeared to interpret their third-stage punishment as being a delayed punishment for free riding in the contribution stage, not a punishment for free riding on the punishing of other low contributions. The above-mentioned evidence in Cinyabuguma *et al*. (2006) that non-punishers receive the least amount of second-order punishment is consistent with our impression that PEO is rare.

These findings suggest that rather than widespread use of higher-order punishment opportunities for the purposes proposed by Henrich and Boyd (PEO) or those suggested by Cinyabuguma *et al*. and Denant-Boemont *et al*. (PEC), the differences in induced cooperation and efficiency observed among our treatments are due mainly to the different patterns of counter-punishment. That is, counter-punishment is more decidedly aimed at perverse than at pro-social first-order punishers in the F treatments than it is in the E treatments.

What accounts for that difference? Concern about the possibility of being punished by third parties for inappropriate punishing behavior may be playing a role despite our failure to detect clear instances of such punishment, since perceived threats needn't be carried out to have an effect. Exposure to more complete information about the overall pattern of punishing in the group in F treatments may also play an important role in its own right. That exposure may help subjects to see an emerging consensus about who it is appropriate to punish.

As for our treatment dimensions other than the ego-centric versus full information distinction, eliminating a separate stage for higher-order punishment seems to have reduced the inefficiency induced by ego-centric counter-punishment opportunities in treatment E2, and contributions are also relatively high in F2. The presence of information on subjects' past play probably helped to raise efficiency in F3h above that in F3n. Even E3h performs better than E3n, despite the fact that the history being shown has an ego-centric bias in it.[15]

[15] Fully disentangling the effects of number of stages, ego-centered vs. full information, and display of history is unfortunately not possible with our data. We estimated regressions in which dummies for each of these three dimensions of treatment variation and their interactions are explanatory variables, with either a subject's average contribution over all 15 periods or her average payoff over those periods as the dependent variable. In OLS regressions, the dummy variable for three rather than two stages obtains a significant negative coefficient, that for display of history a significant positive coefficient, and the interaction between the three-stage dummy and the full information dummy a significant positive coefficient. However, individuals' behaviors are not independent of other individuals in their groups, coefficients lose their significance when errors are clustered by group, and regressions using group level observations also fail to generate significant coefficients, which may be attributed in part to the fewness of observations. The OLS results can still be viewed as suggestive; see Appendix Table B.6.

Finally, it may be recalled that Denant-Boemont *et al*.'s "full information" treatment, most similar to F3n among those conducted by us, yielded contributions significantly lower than those in their basic treatment (which resembles our Reference), although higher than those in their "revenge only" treatment. Why did full information on first-order punishments fail to promote cooperation for Denant-Boemont *et al*.? The most important difference of their "full information" treatment from F3n is that in it, individual identification is scrambled after each period, whereas our subjects' identifications and screen positions remain fixed. This, in principle, makes informed third and higher order punishment possible in F3n. In F3h, especially, information on punishment and contributions remains easy to reference throughout the experiment, making it possible to take into account when choosing punishments many periods later. The performance ordering of our F3h and F3n and Denant-Boemont *et al*.'s "full information" treatments thus further supports the theme that fuller information and more complete freedom to engage in higher-order punishment aids, rather than undermining, the achievement of voluntary collective action.

4. Conclusions

In the recent experimental economics literature, the question has been raised of whether the apparent salutary effects of permitting peer-to-peer sanctions in some public goods experiments may fail to be robust to permitting realistic identification and counter-punishment of punishers. The theoretical literature on the evolution of cooperation has in contrast emphasized the potential importance of the higher-order punishment of those failing to punish norm-violators for the emergence and stability of voluntary cooperation. We designed experiments to further investigate whether opportunities to punish others based on their first-order punishing decisions are helpful or harmful to cooperation. In a treatment closely resembling that of Nikiforakis (2008) and its replication by Denant-Boemont *et al*. (2007), we confirm that when subjects are shown information only about the amount of punishment they themselves receive from identifiable others and have a

dedicated opportunity to punish back in a salient format, the addition of these elements to the original cooperation-and-punishment design has a seriously deleterious effect on cooperation. But removal of the dedicated counter-punishment stage (forcing retaliation to wait until the next period), or provision of more depth of historical information, even though still ego-centric, prove sufficient in our settings to eliminate the negative effect of counter-punishment. And when subjects are provided with more general higher-order punishment opportunities, including but not limited to counter-punishment, efficiency is greater than that in a simple Reference treatment with no higher-order punishment (apart from any "blind revenge" that may take place). The improvement in contributions and efficiency was statistically significant when subjects were provided with broad information on the history of past decisions and had a second punishment opportunity in each period.

Our results suggest that the shielding of subjects from counter-punishment in first generation contribution and punishment experiments was *not* crucial to achieving higher and more sustained levels of cooperation. Although it may indeed be more realistic to think of situations in which peer-to-peer punishment can lead to counter-punishments and while this might well make some punishers think twice, there is also likely to be some observability of punishment by third parties, and norms can emerge wherein most group members understand that punishment of free-riders is generally applauded whereas punishment of cooperators is frowned upon. Full information on who punished whom combined with symmetric opportunities to engage in higher-order punishment and ongoing identifiability of individual group members actually aids, rather than undermining, the cooperation enhancing effects of informal sanctions.

Acknowledgments: We thank Jacob Murray and Iñaki Arbeloa for their help preparing and conducting the experiments. Pilot treatments in collaboration with Jean-Robert Tyran helped us launch this research. The Department of Economics at Brown University provided funding.

References

Axelrod, R. 1986. An Evolutionary Approach to Norms. American Political Science Review 80, 1095-1111.

Bochet, O., Page, T., Putterman, L., 2006. Communication and Punishment in Voluntary Contribution Experiments. Journal of Economic Behavior and Organization 60, 11-26.

Carpenter, J., 2007. Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods. Games and Economic Behavior 60, 31-51.

Cinyabuguma, M., Page T., Putterman, L., 2004. On Perverse and Second-Order Punishment in Public Goods Experiments with Decentralized Sanctions, Working Paper 2004-12, Brown University Department of Economics.

Cinyabuguma, M., Page, T., Putterman, L., 2006. Can Second-Order Punishment Deter Perverse Punishment? Experimental Economics 9, 265-279.

Denant-Boemont, L., Masclet, D., Noussair, C. N., 2007. Punishment, Counter-punishment and Sanction Enforcement in a Social Dilemma Experiment. Economic Theory 33, 145-167.

Engel, C., Kube, S., Kurschilgen, M., 2011. Can We Manage First Impressions in Cooperation Problems? An Experimental Study on "Broken (and Fixed) Windows." Max Planck Institute for Research on Collective Goods, Bonn, Germany.

Falk, A., Fehr, E., Fischbacher, U., 2005. Driving Forces Behind Informal Sanctions, Econometrica 73, 2017-2030.

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

Fehr, E., Gächter, S., 2002. Altruistic Punishment in Humans. Nature 415, 137-140.

Henrich, J., 2004. Cultural Group Selection, Coevolutionary Processes and Large-scale Cooperation. Journal of Economic Behavior and Organization 53, 3-35.

Henrich, J., Boyd, R., 2001. Why People Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas. Journal of Theoretical Biology 208, 79–89.

Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial Punishment Across Societies. Science 319, 1362-1367.

Locke, J. 2005. [1739]. Two Treatises of Government and a Letter Concerning Toleration. Digireads.com Publishing, Stilwell.

Masclet, D., Noussair, C., Tucker, S., Villeval, M.-C., 2003. Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. American Economic Review 93, 366-380.

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

Nikiforakis, N., Normann, H.-T., 2008. A Comparative Statics Analysis of Punishment in Public Goods Experiments. Experimental Economics 11, 358-369.

Nikiforakis, N., Noussair, C., Wilkening, T., forthcoming, "Normative Conflict and Feuds: The Limits of Self-Enforcement," Journal of Public Economics (in press).

Nicklisch, A., Wolff, I., 2011. Cooperation Norms in Multiple Stage Punishment. Journal of Public Economic Theory 13, 791-827.

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

Page, T., Putterman, L., Unel, B., 2005. Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency. Economic Journal 115, 1032-1053.

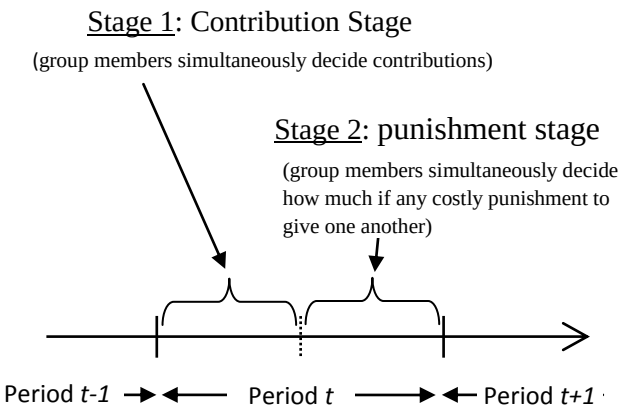
Sefton, M., Shupp, R., Walker, J., 2007. The Effect of Rewards and Sanctions in Provision of Public Goods. Economic Inquiry 45, 671–690.

Sober, E., Wilson, D.S., 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press.

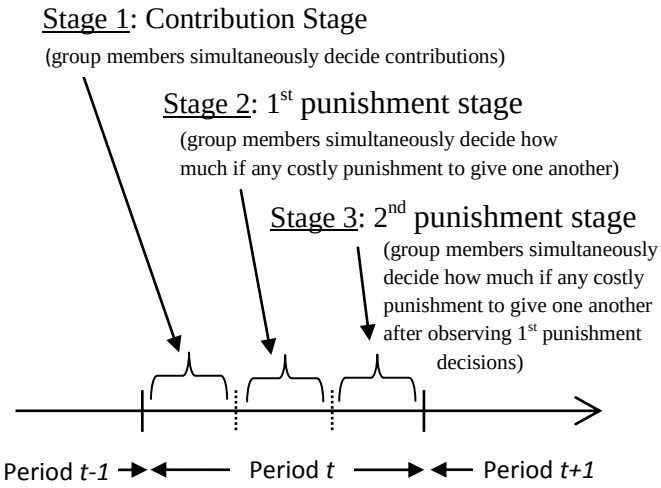
Zelmer, J., 2003. Linear Public Goods Experiments: A Meta-Analysis. Experimental Economics 6, 299-310.

Zizzo, D.J., 2010. Experimenter Demand Effects in Economic Experiments. Experimental Economics 13, 75-98.

Fig. 1. Temporal structure of each period



(a) Two Stage Periods (Reference, E2, F2)



(b) Three Stage Periods (E3n, E3h, F3n, F3h)

Fig. 2. The trends of average contribution to the public account

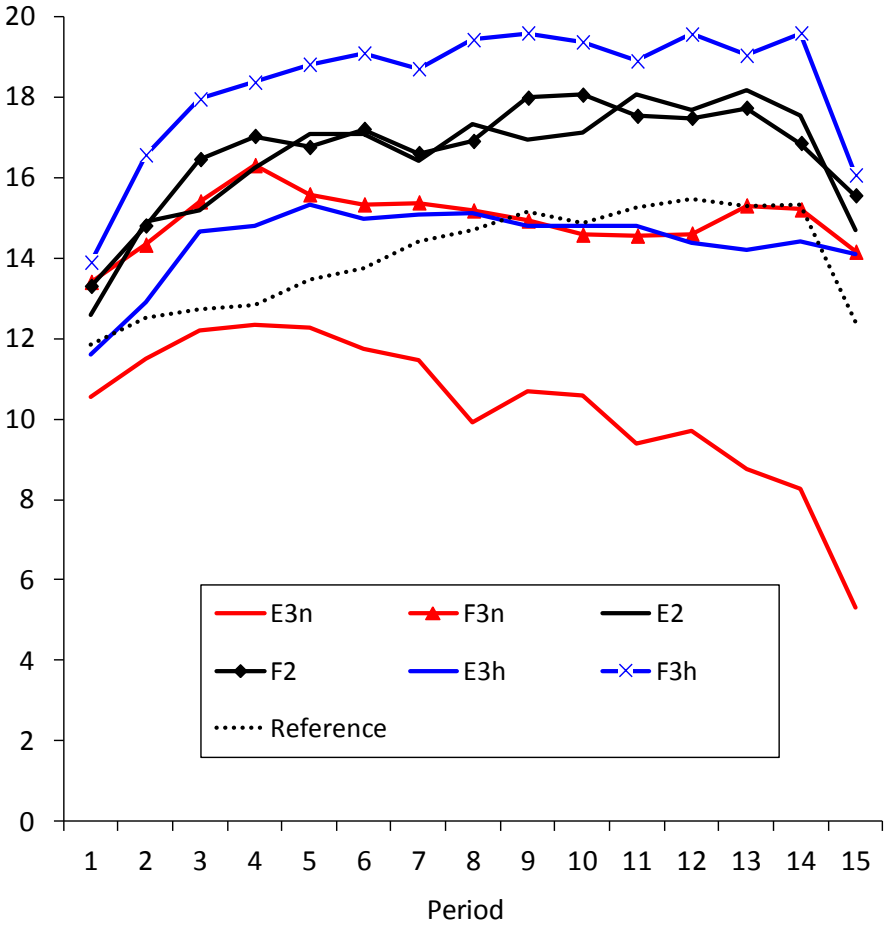


Table 1. Summary of treatments, sessions, and subjects

Treatment	Information Structure[1]	The number of stages in each period	History[2]	higher order punishment opportunities	Total number of sessions	Total number of groups	Total number of subjects
Reference	N	2	n	NO	2	10	40
E3n	E	3	n	YES	2	10	40
F3n	F	3	n	YES	2	10	40
E2	E	2	h	YES	2	10	40
F2	F	2	h	YES	2	10	40
E3h	E	3	h	YES	2	9	36
F3h	F	3	h	YES	2	9	36
Experiment as a whole					14	68	272

Notes: [1]N = no information on who punished whom, E = "Ego-centered information," F = "Full information."
[2]n = "no history of past periods' punishment shown," and h = "history of past periods' punishment shown."

Table 2. Information Available to Subjects in each Treatment

Treatment	Stage 1: Contribution Stage in Period *t*	Stage 2: First Punishment Stage in Period *t*	Stage 3: Second Punishment Stage in Period *t*
Reference	No Information	Contribution decisions[1] in period *t*	N.A.
E3n	No Information	Contribution decisions in period *t*	Stage 2 punishment decisions of group members who have punished you in period *t*
F3n	No Information	Contribution decisions in period *t*	Stage 2 punishment decisions of all group members in period *t*
E2	No Information	(1) Contribution decisions in period *t* (2) Contribution and punishment decisions of those who have punished you in period *t-1*	N.A.
F2	No Information	(1) Contribution decisions in period *t* (2) Contribution and punishment decisions in period *t-1* and average up to period *t-2* of each group member	N.A.
E3h	No Information	(1) Contribution decisions in period *t* (2) Contribution and punishment decisions in period *t*-1 and average up to period *t-2* of each of those who have punished you	(1) Contribution decisions in period *t* (2) Contribution and punishment decisions in period *t* and average up to period *t-1* of each of those who have punished you in period *t*
F3h	No Information	(1) Contribution decisions in period *t* (2) Contribution and punishment decisions in period *t*-1 and average up to period *t-2* of each group member	(1) Contribution decisions in period *t* (2) Contribution and punishment decisions in period *t* and average up to period *t-1* of all members

Note: In each treatment, shows separately the amount contributed by each group member.

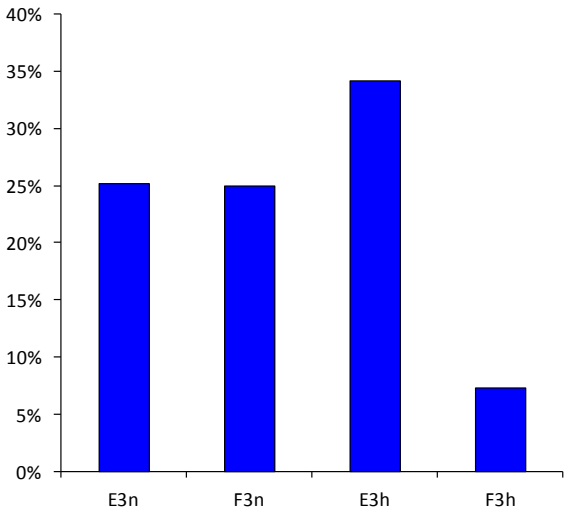
Table 3. Determinants of punishment received in Stage 2 of treatments E2 and F2, including motives for higher-order punishment

Dependent variable: total punishment received by subject *j* in Stage 2 in Period *t*

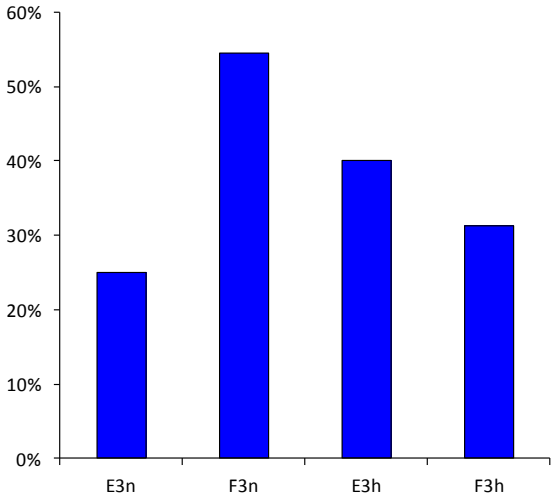
Independent Variable	E2 [1]	F2 [2]
Average Contribution in Period *t*	-0.036 (0.021)	-.011*** (0.025)
Absolute Negative Deviation in Period *t*	0.37*** (0.095)	0.36*** (0.11)
Positive Deviation in Period *t*	0.014 (0.030)	0.013 (0.016)
(a) Total 2nd Stage punishment subject *j* gave to below average contributors in period *t-1*	0.16** (0.052)	-.022 (0.044)
(b) Total 2nd Stage punishment subject *j* gave to above or equal to average contributors in period *t-1*	-0.079 (0.073)	0.26** (0.082)
Constant	0.61 (0.39)	1.91** (0.51)
# of Observations	560	560
F	19.93	40.71
Prob > F	.000	.000
R-Squared	.4641	.3467

Notes: Individual fixed effect linear regression with standard errors clustered by group. The first three explanatory variables explain punishment as a function of period *t* contribution following the specification of Fehr and Gächter (2000). Positive deviation is *j*'s contribution minus the average contribution of other group members, if that difference is positive, otherwise 0. Absolute negative deviation is the average of others' contribution minus *j*'s contribution, if that difference is positive, otherwise 0. A significant negative coefficient on absolute negative deviation is also reported in Fehr and Gächter (2000) and other similar studies. Our specification as a whole allows Stage 2 punishment in period $t > 1$ to be conditioned on both Stage 1 contribution in *t* and Stage 2 punishment in $t - 1$. Observations referencing punishment received in period 1 are omitted due to absence of previous period information.

Fig.3. 3rd Stage counter-punishment as proportion of 2nd stage punishment events



(a) The percentage of events in which the second stage pro-social punishers received counter- punishment in Stage 3 out of the total number of pro-social punishment events in Stage 2[#1]



(b) The percentage of events in which the second stage perverse punishers received counter- punishment in Stage 3 out of the total number of perverse punishment events in Stage 2[#2]

Notes: [#1] The punishment is "pro-social" if it is directed to those who contributed less than the average contribution in their group. [#2] The punishment is "perverse" if it is directed to those who contributed more than or equal to the average contribution in their group.

Table 4. Determinants of higher-order punishment received in E3n, E3h, F3n and F3h

Dependent variable: total punishment received by subject *j* in Stage 3 in Period *t*

Independent variable	E3n (1)	E3h (2)	F3n (3)	F3h (4)
(a) Total 2nd Stage punishment subject *j* gave to below average contributors	0.11 (0.11)	0.27* (0.12)	0.13 (0.073)	0.013 (0.029)
(b) Total 2nd Stage punishment subject *j* gave to above or equal to average contributors	0.027 (0.023)	2.70*** (0.14)	0.24*** (0.040)	0.36** (0.14)
Constant	0.37 (0.27)	-0.18*** (0.31)	0.45*** (0.012)	0.11*** (0.0078)
# of Observations	117	64	600	540
F	.72	---	20.28	3.59
Prob > F	0.5143	---	0.001	0.077
R-Squared	0.1194	0.1899	0.027	0.042
F Test on (a) = (b)				
F	0.86	88.20	3.61	5.91
p-value	.3784	.000	.0898	.0411

Notes: Individual fixed effect linear regression with standard errors clustered by group. In columns (1) and (2), only observations in which subject *j* gave a positive amount of Stage 2 punishment to at least one subject in his or her group are used, since no 3rd stage punishment opportunities are available otherwise.
** and *** indicate significance at the 0.05 level and at the .01 level, respectively.