In Broad Daylight: Fuller Information and Higher-Order Punishment Opportunities Can 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. Varying treatments of a voluntary contributions experiment, we find that availability of higher-order punishment opportunities increases cooperation and efficiency when subjects have full information on the pattern of punishing and its history, when any subject can punish any other, and when the numbers of punishment and of contribution stages are not too unequal.

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.
- Most treatments also permit higher-order punishment.
- In most cases, higher-order punishing opportunities do not harm cooperation.
- When subjects know only who punished them, designated opportunities to counter-punish only are harmful to cooperation and efficiency.
- More symmetric higher-order punishing opportunities with fuller information and history display aid cooperation and efficiency.

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 one question that remains unsettled is whether the benefits of offering punishment opportunities can survive the possibilities of counter-punishment, feuds and vendettas which arise when individuals know who punished them and have opportunities to punish back. In this paper, we contribute to the discussion of punishment's role in supporting cooperation by reporting a set of experiments that make available both the information and the opportunities with which to counter-punish, and in some treatments also to engage in higher-order punishment more broadly.

Views on the role of higher-order punishment cover a wide range, beginning with theoretical suggestions that such punishment plays a role in supporting the punishing of free riding comparable to the role of the latter in supporting (first-order) cooperation itself. Specifically, a proposal regarding the ultimate (as opposed to proximate) source 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 individual 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

[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). Forces ultimately responsible for a propensity to punish should neither be confused with nor seen as being in competition with proximate

(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 was at least at one time helpful, and perhaps even necessary, for fostering cooperation.

On the other end of the spectrum are recent papers by experimental economists which suggest that higher-order punishments are problematic since they often take the form of retaliation and lead to feuds or vendettas. This 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 dangerous, and frequently destructive, to those who attempt it," and that people therefore willingly cede their rights to individually punish 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), Nikiforakis (2008), Hopfensitz and Reuben (2009), and Engel *et al.* (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] Nikiforakis and Engelmann (2011), Nicklisch and Wolff

---

forces, such as anger, predispositions towards which the ultimate forces may have helped to put in place. Some more recent theoretical papers, for example Janssen and Bushman (2010), reach different conclusions about the evolutionary role of higher-order punishment.

[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

(2011), Nikiforakis, Noussair and Wilkening (forthcoming), and Bolle, Tan and Zizzo (2010) further investigate how feuds can be resource-destroying when retaliation is permitted to go on for many rounds.

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 refer to the (second-order) punishment of (first-order) *non-punishers* as *punishment enforcement for omission* (PEO—i.e., for omitting to punish) and to the (second-order) punishment of (first-order) anti-social punishers or perverse as *punishment enforcement for commission* (PEC—i.e., for committing an unjustified act of punishment).

We investigate whether opportunities to engage in higher-order punishment are beneficial or harmful to cooperation and efficiency by conducting a series of experiments in which we vary 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 (in our case, 15 period) 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

---

affected only by the relationship between the recipient's contribution and the overall pattern of contributions in the group. Both sets of researchers agree that in practice the large majority of cases satisfying the definition of "perverse punishment" would also be classified as "anti-social punishment."

who punished them, and subject 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 four 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, during the next punishment stage, full information about all (first-order) punishments in the group and opportunities to punish any group member one wishes to. In keeping with the spirit of full information, the latter treatments eschew the scrambling of subject identifiers, instead assigning group members subject numbers that remain fixed over all periods of play so that individuals' behaviors can be tracked over time. To disentangle the effects of this feature, we also conduct a fixed ID reference treatment with only one punishment stage. Whether IDs are fixed or changing is our second dimension of treatment variation.

Our third 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. Notice that feuding, in the sense of multiple rounds of retaliation, is possible in our design, but must take place across periods, with contribution decisions sandwiched between rounds of punishment. This difference from designs such as that of Nikiforakis and Engelmann (2011) may be important to differences of outcomes, as discussed further, below.

Finally, while the fixed ID feature of our full information treatments makes it possible in principle to predicate future period punishments on past period actions of

other group members, including punishing, we further vary our treatments with respect to whether information regarding past punishments and contributions is explicitly displayed in later periods. If displayed, this history information serves as a unique source of information about second-order punishing and as a convenient reminder of punishment patterns more generally in the full information treatments. In the ego-centric treatments, any identifiable reference to actions in past periods would be otherwise unavailable, even with perfect recall. With one exception, detailed below, the displayed history covers all play thus far.

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 associated with significantly higher contributions and earnings. The treatment yielding clearly higher contributions and earnings than its corresponding reference treatment is one with symmetric (full) 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 are associated in our experiment with 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, that is for punishing high contributors) of the kind discussed by Denant-Boemont *et al*. Instead of these kinds of "sanction enforcement," we see a more pro-social pattern of counter-punishment itself, including significantly less counter-punishing of first-order

punishment aimed at low contributors than of that aimed at high ones, as the main factor differentiating treatments in which punishment information and higher-order punishment opportunities are symmetric from those in which they are ego-centric. While counter-punishment is equally available to subjects in both ego-centric and full information treatments, then, it seems that common knowledge of the pattern of first-order punishing—the "sunshine" of more complete information—plays a role in fostering cooperation through its impact on the pattern of counter-punishment.

In concluding, we note that our results might not extend to situations in which the primary collective action dilemma can be paused indefinitely while a feud takes its course. We suggest, however, that the interleaving of punishing and contributing decisions embodied by our design may correspond well with some real world settings.

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, *inter alia*), 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. The possibility of retaliation is further reduced by the fact that subject identifiers are scrambled after each period. Subject *i*'s earnings in period *t* are given by

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

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_{jit} is the number of units of punishment subject *j* gives to subject *i* in period *t*. Here, the term in brackets is subject *i*'s earnings from the allocation stage, the second term, $\beta \cdot \sum_{j=1, j\neq i}^{n} p_{jit}$, is her loss due to receiving punishment from other members (with β being the loss to the targeted individual per point of punishment given[3]), and the third term, $\sum_{j=1, j\neq i}^{n} p_{ijt}$, 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 (denoted "*T*") of finitely-repeated play, private payoff-maximizing behavior entails $p_{jiT} = 0$ 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

[3] A fixed ratio of punishment cost to punishment effect is assumed in some but not all of the papers mentioned; see note 6, below.

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. Since proof of these suppositions is unavailable to the decision-maker, such instances are labeled "blind revenge" following Ostrom *et al* (1992). Cinyabuguma *et al*. (2006) and Herrmann *et al*. (2008) find evidence consistent with the conjecture that this is the main cause of observed perverse and anti-social punishment. One might thus anticipate that 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 both conjectures.

But some of the experiments just mentioned have their own restrictions. In treatments such as the counter-punishment treatment in Nikiforakis (2008), its replication by Engel *et al*. (2011), the "revenge only" treatment in Denant-Boemont *et al*. (2007), and what is essentially a replication treatment in the current paper (EGO3—see below), subjects learn only of punishments directed at them, and in the additional stage decide only how much to punish back. This, in our view, runs the risk of engendering an "experimenter demand effect" (Zizzo, 2010). 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 the 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—a possibly problematic observation for the Henrich-Boyd theory. Although first-order punishment of above-average contributors 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.

Denant-Boemont *et al*. (2007) conduct a treatment 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 that treatment, 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 (up to six in total) are available each period, and Nikiforakis and Engelmann (2011) and Nicklisch and Wolff (2011) study similar treatments in which the number of higher-order punishment stages is determined endogenously in each period. Nikiforakis, Noussair and Wilkening (forthcoming) introduce similar punishment opportunities in a treatment that increases the likelihood of feuding by making returns from the public good heterogeneous.

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 might be 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 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 both kinds of punishment enforcement (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 her past punishment actions are displayed on a later-appearing screen along with other information about a subject, despite absence of a fixed position or other identifier). 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. Equal observability of punishment interactions between any two group members (*i* and *j*, *j* and *k*, etc.) may be viewed as lying at the other end of that continuum, with many real-world situations lying in between.[4] 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, as with the designated counter-punishment stage introduced in Nikiforakis (2008). In real world interactions, however, punishment chains reaching up to very high orders may take place. Allowing other activity to pause until subjects have engaged in as many rounds of punishment as they wish to provided that they still have resources, is the way of modeling

[4] 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 and to observe others' displays of 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.

this adopted in Nikiforakis and Engelmann (2011) and other studies of feuding. We suggest an alternative approach. We think of individuals as allocating resources to two activities—contributing (or not) to the public good, and punishing (or not punishing) others. An example would be a work group operating under a joint incentive such as profit-sharing, in which any peer punishment meted out in the form of snubbing, bad-mouthing, etc., takes place in a context in which work remains more or less on-going. We worry that giving subjects more opportunities for punishing than for contributing might induce an experimenter demand effect.[5] Having exactly one extra stage—the third stage of the period—available for higher-order punishment—as in the counter-punishment ("revenge only") treatments of Nikiforakis (2008) and Denant-Boemont *et al*. (2007), as well as in the latter's "no revenge" and "full information" treatments—is a possible 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. Both permit feuds to unfold over many rounds, but each requires what is arguably the primary activity under study, that of allocating or not allocating resources to the public good, to be the focus of attention in at least one out of every three decision stages.

[5] 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 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). Nevertheless, the experimental design dictates that a fresh punishment opportunity follows each previous such opportunity in which at least one group member punishes, until resources are depleted. The possibility of imbalance between attention to punishing decisions and attention to contributing decisions may therefore still be an issue.

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. Because identifiers are not scrambled in three of those six treatments, we add a second reference treatment that has no higher-order punishment opportunities but keeps common identifiers across periods. The eight treatments are 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 informed that the experiment will be over at the end of 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*.[6] The punisher pays one point to reduce the earnings of the targeted individual by three points ($\beta = 3$).[7] To avoid the possibility that subjects have to pay the experimenter for losses, we constrain earnings net 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

[6] 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. As noted by Casari (2005), 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), Nikiforakis and Normann (2008), and papers on higher-order punishment such as Nikiforakis and Engelmann (2011).

[7] 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. Nikiforakis and Engelmann (2011) use a 1:2 ratio.

themselves chose to impose. Net losses, in practice rare and limited to a few periods, are covered out of earnings from other periods.[8] Earnings in a given period are accordingly:

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

where p_{jit} is the number of units of punishment subject *j* gives to subject *i* in the first punishment stage in period *t*, pp_{jit} is subject *j*'s punishment of subject *i* due to *i*'s punishing behavior, and S_{jt} is the set of subjects about whom subject *j* is provided with information on which to base higher-order punishment.

In three treatments, dubbed EGO2, EGO3 and EGO3hist, each subject learns following a period's first punishment stage only who punished himself or herself by how many points, while in three counterpart treatments, dubbed FULL2, FULL3 and FULL3hist, subjects learn the amounts of all bilateral first-stage punishments within the group. Four treatments—EGO3, EGO3hist, FULL3 and FULL3hist—add a second opportunity to punish each period, hence they have three stage periods (contribution, first punishment stage, second punishment stage). Two treatments—EGO2 and FULL2—do not add an extra stage. In EGO2, 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. FULL2 is set up similarly, but in it subjects have information about all group members' punishments of one another, so the punishment stage can take into account not only

[8] The constraint that first-stage earnings minus punishment received cannot fall below zero was binding in 23 out of 4,860 periods of individual subject play in the eight 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. Ruling out large negative earnings due to punishment is similarly implemented in Fehr and Gächter (2000) (who constrain effective punishment points received to 10 even if the sum of points given exceeds that number), Fehr and Gächter (2002), Denant-Boemont *et al.*, Bochet *et al*. and Page *et al*. Nikiforakis and Engelmann's rule that a subject cannot be punished further if her net earnings of the period are already zero serves a similar purpose. While we cannot rule out unanticipated effects, our sense is that the constraint has little impact on overall behaviors given the fewness of cases in which it is binding.

others' contributions and punishment of oneself but also punishment of others, which makes each type of punishment enforcement—for omission (PEO) and for commission (PEC)—possible. Also, in FULL2, but not EGO2, subjects are shown average past contributions and punishments going back to the start of play, rather than information about the most recent past period only.

Following Fehr and Gächter (2000), Nikiforakis (2008), and other papers, subject identifiers are scrambled after each period in the three EGO treatments and in the corresponding benchmark treatment, Reference (random ID). Counter-punishment is facilitated in EGO2 by a display of information about each group member's most recent past punishment of oneself. In contrast, but like Nikiforakis (2008), Denant-Boemont *et al*. (2007) and Engel (2011), EGO3 has a third stage in each period for counter-punishment, although no individually-linked information survives beyond the current period. EGO3hist has three stages per period, like EGO3, but provides information on who punished one by how much in all past periods thus far, on average, opening the door to long-running feuds.

Because the spirit of the FULL treatments is one of full information, subject identifiers and screen positions are not scrambled from period to period in them. Reference (fixed ID) provides a comparable baseline having unscrambled IDs but no informed higher-order punishment possibilities. FULL3 has a dedicated stage for higher-order punishment of any group member based on full information on the pattern of punishments, paralleling the FULL treatment in Denant-Boemont *et al*., but subjects do not learn who specifically punished them and by how much, in that stage. In FULL3hist, by contrast, subjects' computer screens display a matrix of information about all past bi-lateral punishments in the group during both punishment stages of a period.[9] As

[9] During the first punishment stage of, say, period *t* in FULL3hist, subjects are shown all combined bi-lateral punishment amounts of period *t* - 1, summing that period's first and second punishment stages. Information about all first stage punishments in period *t* itself is then separately displayed during the second punishment stage (stage 3) of period *t*. Stage 3 punishment amounts of, say, period *t*, can thus be inferred by subtraction of last period's stage 2 punishment amounts from the summed amounts displayed during the first punishment stage of period *t* + 1. On the right hand side of screens in both punishment

mentioned above, FULL2 is a two-stage treatment in which higher-order punishment can take place simultaneously with punishment conditioned on contribution decisions. Unlike EGO2, in FULL2 higher-order punishment is not restricted to counter-punishment but can include PEO and PEC. A matrix of information about all punishments since the first period of play is available in FULL2, as in FULL3hist (whereas in EGO2, information on past play concerns the most recent period only).

In treatments having three stages per period, we can think of the p_{ijt} terms of Eq. (2) as indicating punishments in stage 2 (the first punishment stage) and the pp_{ijt} 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 in the regression analysis of Section 3.[10]

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.

2.3 A Behavioral Predictive Framework

It is well known that standard economic theory predicts no contributions and no punishments in the one-shot or finitely-repeated VCM with opportunities to punish. The same logic applies to higher-order punishment. We devote this brief section to informally discussing some behavioral considerations that are helpful to organizing our findings. The guiding idea is that of subject heterogeneity with respect to the

---

stages, subjects see information about the average contribution and average summed (stage 2 and stage 3) bi-lateral punishments of all periods prior to the most recent one.

[10] Perfect identification of punishments given as first or second-order is in principle impossible also in the FULL3 and FULL3hist treatments, since it cannot be ruled out that a given punishment in a later period is partly in response to punishments of an earlier one. The possibility of multi-period feuding is desirable, in our view, both for purposes of realism and because as it leaves the door open to greater comparability with the feuding experiments mentioned earlier.

predisposition to cooperate provided that others do so (Fischbacher, Gächter and Fehr, 2001; Page, Putterman and Unel, 2005; Fischbacher and Gächter, 2010).

Assume that in the subject pool one finds a large number of individuals who are willing to cooperate with others by contributing to the public good provided that others do the same, and that within this group some conditional cooperators are sufficiently angered by others' free-riding (contributing less) that they prefer to sacrifice some earnings to impose still greater losses on them. Further assume that conditional cooperators consider any punishment they receive—whether for contributing or for punishing free riders—to be unjustified, yielding a willingness on their parts to punish back such punishers. Some of them may even feel strongly enough about the punishment of cooperators that they will punish group members who punish a high contributor other than themselves (that is, engage in PEC). Also assume that there are a smaller number of individuals unwilling to contribute even if others do so, and that individuals of this type consider being punished by others as unjustified and are willing to punish back if punished. Lastly, assume that some individuals, whom we'll call 'neutral,' are not strongly predisposed in either direction, and that whether such individuals fit the conditionally cooperative or the free riding pattern depends on how others are observed to behave, due to a tendency to conform to group norms of behavior.[11] In the VCM without punishment, some conditional cooperators initially test the waters by contributing, but finding free-riding by others, they gradually reduce their contributions. If the proportion of conditional cooperators is sufficiently large, adding a single punishment stage without revealing who punished whom may transform the dynamic by giving cooperators opportunities to punish free riders and to bring the contributions of the latter and of the neutral subjects' up instead of bringing their own contributions down.

[11] In Fischbacher *et al*., there is an additional type, labeled hump-shaped or triangle contributors, from which we abstract here because we feel it to be more important to emphasize the presence of relatively neutral subjects not strongly wedded to either of the polar types. Indeed, a more realistic description of many populations might allow for a large proportion of individuals who may be nudged from one orientation to another by contextual cues and others' behaviors. In a sense, triangle contributors are conditional cooperators as long as they anticipate fairly modest cooperation rates by others, but gradually become free riders as they perceive others' cooperation rates to be approaching the maximum possible.

Some free riders attempt to punish back by blindly punishing high contributors, with the extent of damage depending on the relative numbers of group members of each type. When, in addition, group members are told who punished them and to what extent and are given a dedicated opportunity to punish back, free riders will use the new stage to punish back and some cooperators will be deterred from punishing, so the availability of punishment will be of less benefit to cooperation.

Suppose instead that subjects receive information about all pairwise punishments in their group. In the typical subject group, the large majority of first-order punishments are aimed at low contributors, so information on the overall punishing pattern encourages neutral types to join the cooperators in contributing to the public good and in some cases even join in punishing free riders.[12] Free riders may retaliate against their punishers, but if identifications are fixed, the latter can now retaliate back, perhaps joined by other cooperators who perceive the initial retaliations as unjustified and as working against group efficiency. The numerical superiority of conditional cooperators and neutral subjects is likely to make for a pattern of contributions even more sustained than when there is only one punishment stage and absence of feedback and identifiability, since both types of "deviant" or anti-social behaviors—free riding, and punishing cooperators—are punished. Costs of punishing will initially lower efficiency, but if a relatively cooperative pattern is established, punishment costs are likely to decline, so earnings may become higher than in the environment permitting only one order of punishment.

The general contours of these remarks can be summarized by saying that adding feedback only on punishment subjects receive and adding only opportunities to punish back is likely to deter punishment and thus cooperation, while adding full information on punishing patterns, their association with contributions, and opportunities to punish any punishment action, with continuity of identification, may help to strengthen cooperation if the share of strongly predisposed free riders is small.

[12] Once an individual has agreed to make the sacrifice of contributing to the group account, she is likely to feel much the same anger at any free riders as do those more strongly disposed towards cooperating from the outset. See Gürerk et al. (2006) for examples of subjects who switch to punishing once they switch to contributing to the public good.

Of course, there are formal social preference models, some of them less psychological in flavor, that predict positive contributions and punishment of free riding under certain parameter values. Providing formal underpinnings and comparing the model predictions for each treatment studied by us may be useful after empirical regularities become clearer through experimentation, but that task lies beyond the scope of our paper. We accordingly turn, without further discussion, to analyzing the results of our experiment.

3. Results

17 experiment sessions were conducted in a computer lab at Brown University between October, 2011 and January, 2013.[13] In total 324 students were recruited from the general undergraduate population, representing concentrations (majors) in the humanities, social sciences, and sciences, with 19.6% being economics concentrators (slightly higher than their share in the general student population) and 52.9% female (also slightly above a representative share).[14] 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 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. The latter averaged $19.58 (1 experimental point = $0.05) plus a $5 show-up fee.

Figure 2 displays the trends of average contributions and earnings period by period for each treatment, with full information and fixed ID Reference treatments in the

[13] We planned 2 sessions for each treatment. After discovering that subjects in three of the groups in the first session of the EGO3hist treatment received some erroneous feedback on their screens due to a programming error, we conducted an additional session of that treatment.

[14] 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."

left panels and ego-centric information and scrambled ID Reference treatments in the right panels. Both Reference treatments display the pattern of contributions already familiar from first-generation contribution and punishment treatments. The average contribution begins around 60% of endowment, and then trends upwards towards 75% of endowment before a last-period decline.[15]

Although every treatment other than the Reference treatments offers subjects the opportunity to engage at least in counter-punishment and possibly in other kinds of higher-order punishment, all but one shows no sign of contributions or earnings being lower than in the corresponding Reference. Average contribution is higher than in Reference (fixed ID) (although not necessarily significantly so) in every period for the FULL2 and FULL3hist treatments, and average contributions are higher in EGO2 than in Reference (random ID) in all but period 1. 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 (fixed ID) only for the treatment having the highest average contribution curve, FULL3hist ($p = .014 < 0.05$, 2-tailed test).[16] Contributions in FULL3hist are also significantly different than those in

[15] Insofar as Fehr and Gächter (2000) scrambled identifiers to prevent reputation formation and insofar as this was partly motivated by a desire to avoid vendettas, it is of interest to check whether there is in fact more punishment of high contributors in Reference (fixed ID) than in Reference (random ID). In the event, contributions and earnings are on average slightly lower in the former than in the latter treatment in the early periods, but the difference is not statistically significant according to a group-level Mann-Whitney test even when only periods 1 – 7 are considered. We do find somewhat more cases of punishment of subjects who have thus far been among their group's highest contributors by low contributors: 29 cases in 249 opportunities in Reference (fixed ID) versus 8 of 201 opportunities in Reference (random ID) during the first seven periods, which supports the conjecture that revenge, somewhat less blind in the fixed ID treatment, may help account for the apparent difference. (Revenge is less blind in the sense that a low contributor, for example, knows who was a high contributor in past periods, and not only in the current one; revenge remains blind, nonetheless, because individuals still never see who punished whom.) The relatively large and significant positive coefficient on the Positive Deviation term in Appendix Table B.5 also indicates that punishing of high contributors was somewhat more common in Reference (fixed ID). Nevertheless, it is clear that whether or not identifiers are scrambled does not change behaviors dramatically. (We could not find other comparisons in the literature of finitely repeated partner VCM treatments with punishment—but without communication or other added elements—that differ only with respect to whether identifiers are or are not scrambled, as do our two Reference treatments.)

[16] For the number of observations in this and other reported Mann-Whitney tests using group-level observations, see the numbers of groups by treatment listed in Table 1.

FULL3 ($p = 0.034 < 0.05$, 2-tailed test). Earnings differences are statistically significant in the same two cases, with the same direction, as shown in Appendix Table B.2.

In one treatment, contributions are clearly *lower* than in the corresponding Reference treatment: EGO3, the ego-centric information treatment modeled on the counter-punishment condition of Nikiforakis (2008) and Denant-Boemont *et al*.'s "revenge only" treatment. The upper right panel of Figure 2 shows EGO3 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 EGO3 than in EGO2 and Reference (random ID) ($p = 0.005$ and 0.082, respectively, 2-tailed test).[17] Average earnings are significantly lower in EGO3 than in the same two treatment (EGO2, and Reference (random ID), $p = 0.082$, and 0.048, respectively) , and earnings are also significantly lower in EGO3 than in EGO3hist ($p = 0.003$, all tests 2-tailed; see again Appendix Table B.2).

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. Looking first at the total amounts of punishing, measured as the sum of costs incurred by punishers and by those targeted for punishment, Table 3 shows considerable variation among treatments in the amounts of first-order (stage 2) punishing: almost four times as much in Reference (Fixed ID) as in FULL3, a treatment in which it was evidently quite common for subjects to delay punishment until stage 3 so as to escape retaliation.[18] The addition of second-order (stage 3) punishing tends to narrow overall differences, to a gap of just 11% in the aforementioned treatments, and a

[17] Statements based on 2-tailed Mann Whitney tests using group-level observations averaged over 15 periods. Contributions are also lower in EGO3 than in the three FULL treatments according to two-tailed Mann-Whitney tests, although it should be born in mind that these treatments differ with regard to persistence of identifiability across periods.

[18] Escaping retaliation is likely because subjects are never shown who gave what stage 3 punishment, in the FULL3 treatment. Subjects could always delay first-order punishing in FULL3 but not in EGO3 because in the latter, a subject could punish in the third stage only a group member who had punished her in the same period's second stage.

maximum gap of 1.9:1 between the Reference (Fixed ID) and the two "hist" treatments. While some of the differences in first-order punishing are statistically significant, no treatment shows a statistically significant difference from any other with respect to overall punishing (see Appendix Table B.3).[19] A broad impression, however, is that especially among the fixed ID treatments availability of higher-order punishment opportunities somewhat discourages first-order punishing, with the difference partly made up by the second punishment stage when available.

The italicized rows in the upper portion of Table 3 give information about the pattern of second-stage (mainly first-order) punishment, while corresponding rows under variable (ii) give information about the pattern within that part of third-stage (second-order) punishment that can be classified as counter-punishing. We first observe that the proportion of punishment expenditure (and loss to those targeted) classifiable as anti-social or perverse is somewhat higher in the EGO3 and EGO3hist treatments than in the FULL3 and FULL3hist treatments, although the differences with respect to numbers of events (the third and fourth italicized rows) run in the opposite direction. Overall, 82.6% of second stage punishment (in cost terms) was targeted at low contributors. The rows under (ii) indicate more consistent differences. Comparing the number of points of counter-punishment returned per point of stage 2 punishment in the 3 stage treatments, we see that counter-punishment strength was three to fifteen times as great when a perverse or anti-social punishment was being counter-punished than when normal or pro-social punishment were being responded to, in FULL3 and FULL3hist. By contrast, there is a weaker point for point response to anti-social or perverse than to pro-social or normal punishment in EGO3 and EGO3hist. Similar patterns apply in the bottom two italicized rows, which show ratios of the *proportion* of anti-social or perverse punishment

[19] In the experiment as a whole, 71.9% of subjects punished at least once, 75.0% were punished at least once, and the average subject punished at least one other subject in 7.7% of periods, in Stage 2. For share of Stage 2 punishment opportunities used, the difference between the Reference (fixed ID) and FULL3 is significant at the 5% level, that between FULL2 and FULL3 is significant at the 10% level, and that between EGO2 and EGO3hist is significant at the 5% level. For total points of punishment given in Stage 2, FULL2 has more than FULL3, significant at the 5% level, and Reference (random ID) has more than EGO3, significant at the 5% level. See Appendix Table B.4.

events counter-punished to the corresponding proportion of pro-social or normal punishment events so responded to. The stronger responses to anti-social and perverse punishment may help to explain the lower expenditures on such punishment in the FULL3 and FULL3hist treatments.

As a next step in studying the targeting of punishment, we estimated regression equations following a specification in Fehr and Gächter (2000). In our data as in theirs, we find that 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.[20] This indicates that while motives to punish free riding may be mixed with motives to counter-punish, especially in the two-stage treatments, the predominant motive of punishers at the first punishment stage seems the same as in other experiments. There is substantial variation in coefficient values on the negative deviation variable, however. They are lowest in FULL3 and EGO3, again suggesting less tendency to punish free riding in Stage 2 when there is a third stage to follow and especially in the absence of displayed history. We note that the average punishment received for contributing one less point than the group's average was nonetheless greater than the 0.6 required to render contributing at least the average privately profitable, except in the FULL3 treatment where, according to the estimated coefficient, a subjects lost 0.48 points to punishment for each point less that she contributed.

To better understand the determinants of higher-order punishment, we estimate regression equations in which potential proximate causes for punishing appear as

[20] To conserve space, the regression results are shown in Online Appendix, Tables B.5. 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 EGO3 treatment. In three of the treatments, Reference (Random ID), FULL3 and EGO3hist, there is a significant although small positive coefficient on the positive deviation term, suggesting that contributing too much *above* the average attracted perverse or anti-social punishment. Önes and Putterman (2007, Table 2) find similar results for some treatments. Only one of these three positive coefficients on the positive deviation variable is statistically significant in alternative estimates using a Random Effects Tobit estimator.

explanatory variables. We begin with the two-stage treatments, in which second-order punishment is of necessity mingled with first-order punishment in a single punishment stage, and our regression analysis attempts to tease these components apart, at least approximately. Table 4 shows estimates of random effects Tobit regressions in which the observations are specific to each pair of subjects, *i* and *j*, in a group.[21] The dependent variable is punishment received by subject *j* from subject *i* in period *t*, and explanatory variables include average contribution in the group, absolute negative deviation of *j*'s from *i*'s contribution (a positive number if $C_j < C_i$, otherwise 0), and positive deviation between the contributions of the pair (a positive number if $C_j > C_i$, otherwise 0). Period *t* values of the contribution deviations are included, as in other studies, to account for the main expected reason for first-order punishment, while period $t - 1$ values are included in two of the specifications because of their potential influence on subject *i*'s reaction to having been punished by *j* in that period. Because in EGO2 subjects are shown the previous period's contributions of those who gave them punishment but not of other group members, inclusion of the lagged contribution variable in the column [2] regression requires that we include only those *i*, *j* pairs such that $p_{ji(t-1)} > 0$, which considerably reduces sample size. In regression [1], we estimate a model for EGO2 that drops the contribution lags, in order to include more observations.[22]

In addition to the contribution variables, the regressions include two variables intended to pick up counter-punishment, one being the amount of punishment *j* gave *i* in period $t - 1$ "pro-socially" (i.e., if $C_{j,t-1} > C_{i,t-1}$, variable (vi)), the other any corresponding "anti-social" punishment of *i* by *j* in that period (variable (vii)). Since subjects in the FULL2 treatment were shown amounts of punishment given last period to group members other than themselves, regression [3], for FULL2, also controls for punishment *j* gave in $t - 1$ to group members $k \neq i$ who contributed less than or the same amount as *i*

[21] We use a Tobit estimator because of the large number of zero values of the dependent variable. We control for temporal structure by including a period term, and we take into account the multiple observations of given individuals by adopting a random effects estimator.

[22] Note that since $p_{ji(t-1)} = 0$ for all observations included in [1] but not in [2], differences in the estimates of the coefficients on the past punishment variables between these two column cannot be attributed to differences in past punishment behavior among those subject-period observations excluded from [2].

(variable (viii)) and to any $k \neq i$ who contributed more than i (variable (ix)). Both regressions include a control for period.[23]

Estimates [1] and [3] support the usual pattern of first-order punishment being significantly larger the larger the negative difference between punisher and punishment target, while the positive deviation term and both lagged deviation terms have insignificant coefficients in all three estimates. Presence of counter-punishment for last period's punishing is also strongly supported by five of the six estimates of variables (vi) and (vii), with the coefficients for counter-punishing anti-social punishment acts (variable (vii)) being in all cases larger than those for counter-punishing pro-social punishment (variable (vi)), a regularity on which we comment more when discussing the three stage treatments below. With respect to punishment enforcement variables (viii) and (ix), the first obtains a significant negative sign, suggesting that the more a subject *j* punished others who were relative low contributors, the less second-order punishment did *j* receive from *i*—perhaps indicating approval of pro-social (or "altruistic" [Fehr and Gächter, 2002]) punishment acts.

Turning to the three stage treatments (EGO3, FULL3, EGO3hist, FULL3hist) and focusing first on counter-punishment (punishing back by the recipient of first-order punishment), we first comment briefly on Figure 3, which shows the fraction of second-stage punishment events ($p_{ij}^t > 0$) that are followed by third-stage counter-punishment from the recipient ($pp_{ji}^t > 0$), with observations distinguished by whether the punishment event which might be counter-punished was itself "pro-social" (panel (a)) or "normal" (panel (b)) versus "anti-social" (panel (a)) or "perverse" (panel (b)). In each panel, the upper, gray bars show the percentage of "anti-social" or "perverse" punishment events that are counter-punished, while the lower, black bars show the percentage of "pro-social" or "normal" punishment events that are counter-punished. The proportion of first-order punishment events that are counter-punished is substantial, in the 20 to 60% range in most cases. Importantly, for almost all treatments, anti-social and perverse first-order

[23] Period 1 observations are excluded since no previous period punishment had taken place.

punishments are met by retaliation in higher proportions of cases than are pro-social and normal (non-perverse) ones, with the differences being greatest for FULL3hist, followed by FULL3.[24] Retaliation against first-order punishers of low contributors is especially low in FULL3hist—the treatment that attains highest efficiency. There, only 7.2% of punishments directed at below-average contributors, or 5.6% of pro-social punishments, are followed by counter-punishment. Apparently, punished free riders usually accept their punishment while punished high contributors retaliate, in FULL3hist, a pattern that is much less evident in the EGO treatments.[25]

In Table 5, we study third-stage punishment received by subject *j* from subject *i* as a function of the second-stage (first-order) punishment given by *j* and *j*'s contribution decision in the same period. In EGO3 and EGO3hist, subjects were only able to engage in third-stage punishment of group members who punished them in the period's second stage and only had information about those members' contributions and second stage punishment when making their third-stage decisions, so only observations of *i*,*j* pairs for whom third-stage punishment was a possibility are included, considerably restricting the sample.[26] In FULL3 and FULL3hist, in contrast, any group member could punish any other and all had information about the contributions and about all bi-lateral second stage punishments when making their third-stage decision, so our regressions include all *i*,*j* pair observations for each period. As with Table 4, we use random effects Tobit specifications, and the set of explanatory variables is similar except that higher-order punishment is assumed to be conditioned on first-order punishment of the present period, and only the deviations between *i*'s and *j*'s contributions in the current period are

[24] The ratios of gray to black bar lengths for given treatments are in fact equal to the "relative frequency of counter-punishing anti-social vs. pro-social punishers" (panel (a)) and "relative frequency of counter-punishing perverse vs. normal punishers" (panel (b)) in the bottom two italicized rows of Table 3.

[25] Although the Table 4 regression estimates suggest that counter-punishing of anti-social punishment was also more common than was counter-punishing of pro-social punishment in the two two-stage treatments, we leave those treatments out of Figure 3's event analysis since with first- and second-order punishment co-mingled in a single stage, it's unclear to what degree a given instance of punishment should be classified as a counter-punishment event.

[26] We refrain from estimating variants of the regressions for the larger universe of observations because in this case, unlike Table 4, second-order punishment has its own stage rather than being comingled with punishment of contribution choices.

included.[27] As with Table 4's regressions, specifications for the FULL treatments allow higher-order punishment to be conditioned also on any first-order punishment *j* gave to third parties (punishment enforcement), while those for the EGO treatments leave out the terms in question (variables (vi) and (vii)) since subjects only learned of punishments aimed at themselves.

While the table shows only a single coefficient to be statistically significant in the regressions for the EGO treatments, there are many significant coefficients in those for the FULL treatments, perhaps partly due to sample size. There are indications, first, of additional or delayed punishing of the period's free riders, in the form of positive significant coefficients on the negative deviation term (variable (ii)) and, for FULL3hist, a significant negative coefficient on positive deviation. Turning to genuinely second-order motives for punishing, we find positive coefficients on both the amount of anti-social and the amount of pro-social first-order punishment received by *i* (variables (iv) and (v) respectively). Both variables are highly significant in the regression for FULL3hist, whereas only the first one is for FULL3. For FULL3hist, the coefficient for counter-punishing pro-social punishers is only about half as large as that for counter-punishing anti-social ones, consistent with our impression from Figure 3 that "unjustly" punished subjects punished back more in this treatment. The absence of a significant coefficient for counter-punishment of anti-social punishers (variable (v)) in the FULL3 regression suggests a weakness of efficiency-promoting counter-punishing in that treatment relative to FULL3hist. The difference may help to explain the relatively lower contributions in FULL3.[28]

[27] Although actions in past periods can also affect this period's punishment decisions especially in the three stage FULL treatments, which have fixed subject IDs, they are of more secondary concern than in the two stage treatments, where higher-order punishment can occur only in the next and later periods. We omit lagged terms to avoid cluttering our specifications.

[28] Interestingly, the ratios of the coefficient for counter-punishing of anti-social punishment (variable (v) of Table 5, (vii) of Table 4) to the coefficient for counter-punishing of pro-social punishment (variable (iv) of Table 5, (vi) of Table 4)—a ratio equaling $3.61/1.70 \approx 2.1$ for FULL3hist, $1.20/0.75 \approx 1.6$ in FULL2, and $1.07/2.01 \approx 0.5$ in FULL3—align perfectly with the contribution trends visible in Figure 1: highest for FULL3hist, lowest (among the three treatments) for FULL3, and intermediate for FULL2. This and other indications suggest that the forcefulness with which anti-social first-order punishment was counter-

As with FULL2 in Table 4, the regressions for the FULL treatments also attempt to pick up second-order punishment predicated on punishment given to others—i.e., punishment enforcement. Both coefficients on punishment given to others who contributed less than *i* are negative, as with the corresponding coefficient in FULL2. The coefficient in the estimate for the FULL3 treatment is statistically significant, again suggesting approval of pro-social punishing.

That the regressions show clear patterns of counter-punishment and, to a lesser degree, of differential punishment enforcement, in the three stage FULL but not in the corresponding EGO treatments, is consistent with the possibility that more effective use of the signaling potential of second-order punishment given more general information, more general opportunities to punish, and perhaps greater normative consensus, helps explain the better contribution outcomes in the FULL than in the EGO treatments. While the possible impact of differences in continuity of subject IDs cannot be ruled out and makes direct comparisons between these treatments somewhat problematic, there are reasons to suppose that factor to be of secondary importance given the absence of significant differences between the two Reference treatments.

We also looked at our three-stage treatments, where the evidence should be clearest, for more direct signs of both PEO (punishment enforcement for omission [i.e., for failure to punish]) and PEC (punishment enforcement for commission [i.e., for perversely or anti-socially punishing]). In FULL3 and FULL3hist, we found cases in which a first-order (second-stage) non-punisher received punishment in the third stage, consistent with PEO. But the large majority of these cases can be explained as delayed first-order punishment.[29] To check for PEC, we identified all cases in which an equal-to-

---

punished, in comparison to counter-punishing of pro-social first-order punishment, is a key to the success of both orders of punishment in raising contributions. The relatively high ratio of the coefficients in question for the EGO2 treatment, in Table 4 (in both estimate [1] and [2]), is also consistent with that treatment's strong contribution performance.

[29] In 91.9 % of the FULL3 and 80.0% of the FULL3hist cases in which a first-order non-punisher received second-order punishment, the targeted individuals were low contributors, and by delaying punishment to stage 3, the punisher eluded counter-punishment, especially in FULL3 which lacks history display. Detailed results are included in Appendix B.1 Supporting Analysis. A form of indirect evidence about

or-above-average contributor was (perversely) punished in the first punishment stage and there was a third group member, not the punished subject, who had an opportunity to punish the group member responsible. Limiting this search to potential third-party punishers who were themselves above-average contributors, we found only five and eleven such opportunities in FULL3 and FULL3hist, respectively. Out of these potential cases, we found actual second-order PEC in only one case, in FULL3hist. So observable PEC is also rare, in our data.

Overall, our 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. Counter-punishment is most decidedly aimed at anti-social as opposed to pro-social first-order punishers, according to our regression evidence, in FULL3hist, the treatment that attains the highest efficiency of those studied. Correspondingly, counter-punishment is less differentially aimed at anti-social punishers (Fig. 3(a)), and least differentially aimed at perverse punishers (Fig. 3(b)), in EGO3, the treatment attaining the lowest efficiency.

Since subjects possess the requisite information and the opportunity to counter-punish in both sets of treatments, what accounts for the difference in the relative strengths

---

what was in fact motivating some subjects to punish individuals who failed to punish in the second stage is provided by looking at how the punishment recipients themselves responded to being punished. Regressions shown in Appendix Table B.10 find evidence that such recipients of third-stage (second-order) punishment responded by raising their contributions, not by engaging in the second- stage (first-order) 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 others' 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. We refrain from drawing overly strong conclusions, however, because Denant-Boemont *et al*., adopting a regression specification which includes as an independent variable the level of *j*'s first-order punishing relative to that of others, report statistically significant coefficients suggesting that those punishing less at that stage received more higher-order punishment, consistent with the spirit of the PEO idea. Nikiforakis and Engelmann (2011) report that 30% of the second-order punishment observed by them went to first-order non-punishers, but it is not clear whether alternative explanations such as delayed first-order punishment might be applicable.

of pro- versus anti-social counter-punishing? Concern about the possibility of being punished by third parties for inappropriate punishing behavior (PEC) may be playing a role despite our failure to detect clear instances of it; after all, perceived threats needn't be carried out in order to have an effect. Exposure to more complete information about the overall pattern of punishing in the group in FULL treatments may also play an important role in its own right. That exposure may help subjects to perceive an emerging consensus about who it is appropriate to punish. The additional identifiability of individual group members might also have interacted with the normative power of the sense of consensus, perhaps by inducing a sense of shame in free-riders and perverse punishers (Bowles and Gintis, 2005; Hopfensitz and Reuben, 2009).[30]

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 EGO2, and contributions are also relatively high in FULL2. The presence of information on subjects' past play probably helps to raise efficiency in FULL3hist above that in FULL3, perhaps in part by strengthening the perception of consensus and in part by increasing the danger of punishment for 'inappropriate' behaviors. Even EGO3hist performs better than EGO3, despite the fact that the history being shown has an ego-centric bias in it.

[30] It may be recalled that Denant-Boemont *et al*.'s "full information" treatment, most similar to FULL3 among those conducted by us, yielded contributions significantly lower than those in their basic treatment (which resembles our Reference (random ID)), although higher than those in their "revenge only" treatment (which resembles our EGO3). Why did full information on first-order punishments fail to promote cooperation more for Denant-Boemont *et al*.? The most important difference of their "full information" treatment from FULL3 is that in it, individual identification is scrambled after each period, whereas our subjects' identifications and screen positions remain fixed. In FULL3hist, 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 FULL3hist and FULL3 and Denant-Boemont *et al*.'s "full information" treatments thus further supports indications that fuller information and more complete freedom to engage in higher-order punishment aids the achievement of voluntary collective action, at least when the number of punishment stages per period is not too large. We are grateful to an anonymous referee for suggesting the idea of an interaction between individual identifiability and feelings of shame.

Unfortunately, definitively disentangling the effects of number of stages, ego-centered vs. full information, display of history, and presence or not of fixed IDs may not be possible with our data. We estimated OLS regressions in which dummies for each of these 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 the regressions, shown in Table 6, the dummy variable for three rather than two stages obtains a significant negative coefficient, that for display of history a weakly significant positive coefficient, and the interaction between the three-stage dummy and the full information dummy a significant positive coefficient in both regressions. Individuals' behaviors are not independent of other individuals in their groups, however. The coefficients lose their significance when errors are clustered by group.[31] The alternative of estimating regressions using group level observations also fails to generate significant coefficients, which may be attributed in part to the small number of observations. We accordingly show OLS results for purposes of rough impression, only.

Finally, to explain the sustaining of contributions and earnings in most of our higher-order punishment treatments and their significant enhancement in the FULL3hist treatment in comparison to the findings of Nikiforakis and Engelmann (2011), it is important to remember that while our design potentially allows for feuds of many rounds, those feuds must take place over the course of multiple periods, each of which begins with a new set of contribution decisions. Subject preoccupation with retaliatory motives is most likely thereby attenuated, which probably both reduces the duration of any feuding and reduces the concern over possible feuds as a disincentive to engaging in first-order punishment.[32]

[31] Since there is only one observation per subject, individual fixed effects are ruled out.
[32] Feuds are in fact much more difficult to identify in our data than in Nikiforakis and Engelmann's, since punishing in later periods could be a response to a large number of potential causes. This problem of identifiability would have affected subjects themselves and would probably have tended to dampen any feuding that took place.

4. Conclusions

In the experimental economics literature of recent years, 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. Some theoretical literature on the evolution of cooperation has in contrast emphasized the potential importance, for the emergence and stability of voluntary cooperation, of the higher-order punishment of those failing to punish norm-violators.

We designed experiments to investigate when opportunities to punish others based on their first-order punishing decisions are helpful vs. harmful to cooperation. In a treatment closely resembling that of Nikiforakis (2008) and its replications by Denant-Boemont *et al*. (2007) and Engel *et al*. (2011), we reconfirm 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 and efficiency. 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 fixed identifiers and with more general higher-order punishment opportunities, including but not limited to counter-punishment, efficiency is greater than that in a treatment with no higher-order punishment. The improvement in contributions and efficiency is statistically significant when subjects are provided with broad information on the history of past decisions and have a second punishment opportunity in each period.

Analysis of punishment patterns suggests that presence of higher-order punishment by third parties is not the cause of the difference in outcomes. Rather, counter-punishment itself seems to be more pro-socially or efficiently organized in

treatments with fuller information and history than in ones with ego-centric information and little history retention. This suggests that the main channel through which treatment differences operate is that of changing subjects' perceptions of what actions are legitimate to punish, and perhaps thereby altering their emotional or normative responses to punishment. Finding ways to verify or disconfirm this interpretation would be a useful direction for future research.

We think it important to note that words like "suggest" appear frequently in our discussion because definitive determination of why *i* or *k* punished *j* and why *j* went on to contribute more or less is elusive. Many of the interpretations offered rest unavoidably on chains of inference, and an almost infinite number of alternative conjectures could be posed and assessed with differing techniques, e.g. regression specifications. As a result, we have greatest confidence in those conclusions that stem from non-parametric comparisons of behaviors and outcomes across treatments.

What our results indicate most clearly, we believe, is 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 than observed when no punishment opportunities at all are unavailable. Although it may indeed be more realistic to think of situations in which peer-to-peer punishment can lead to counter-punishments, as authors beginning with Nikiforakis have helpfully pointed out, 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, appears to aid, rather than undermine, the cooperation-enhancing effects of informal sanctions, in our experiment.

To be sure, unlike some experiments on feuding, our subjects are not permitted to pause decision-making on the first-stage collective action problem for large numbers of

uninterrupted retaliation rounds. Which approach is more realistic—permitting uninterrupted feuds or having contribution and punishment opportunities be interspersed, as in our design—depends on the specifics of the environment being modeled. We leave it to readers, and to future research, to judge the extent to which our findings effectively allay concerns about counter-punishment and feuding as obstacles to voluntary collective action.

Acknowledgments: We thank Jacob Murray, Iñaki Arbeloa and Yunan Ji 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. We thank two anonymous referees, an associate editor and the editor whose suggestions helped us to improve the paper substantially.

References

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

Bolle, F., Tan, J.H.W., Zizzo, D.J., 2010. Vendettas. University of Nottingham CeDEx Discussion Paper 2010-02.

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

Bowles, S. and Gintis, H., 2005. Pro-social Emotions. Pp. 337 – 67 in L. Blume and S. Durlauf, eds., The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow. Oxford: Oxford University Press.

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

Casari, M., 2005. On the Design of Peer Punishment Experiments. Experimental Economics 8, 107-115.

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.

Fischbacher, Urs and Simon Gächter, 2010, "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments," American Economic Review 100(1): 541-56.

Fischbacher ,U., Gächter, S. and Fehr, E., 2001. Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. Economics Letters 71, 397-404.

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

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.

Hopfensitz, A. and Reuben, E., 2009. The Importance of Emotions for the Effectiveness of Social Punishment. Economic Journal 119, 1534-59.

Janssen, M. and Bushman, C., 2008. Evolution of Cooperation and Altruistic Punishment When Retaliation is Possible. Journal of Theoretical Biology 254, 541-45.

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.

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

Nikiforakis, N., 2008. Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves? Journal of Public Economics 92, 91-112.

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

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).

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.

Ostrom, E., Walker, J. and Gardner, R., 1992. Covenants with and without a Sword: Self Governance is Possible. American Political Science Review 86, 404-416.

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

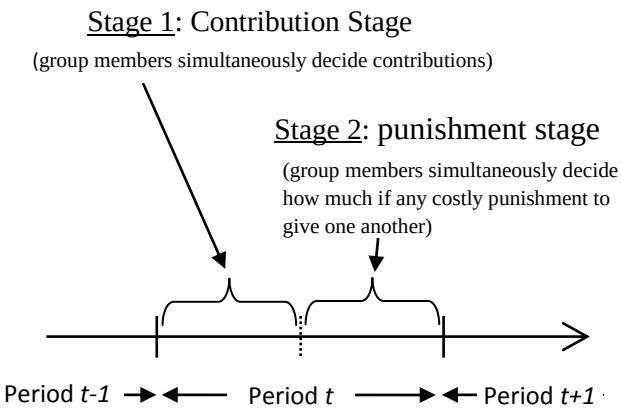
Sefton, M., Shupp, R., Walker, J., 2007. The Effect of Rewards and Sanctions in Provision of Public Goods. Economic Inquiry 45, 671–690.

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

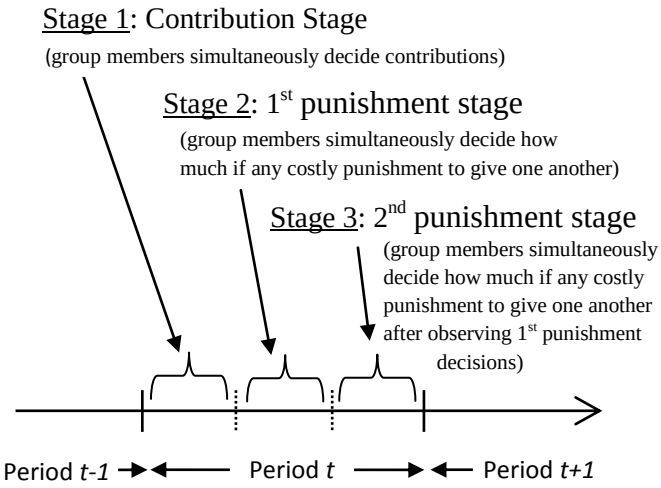
Zelmer, J., 2003. Linear Public Goods Experiments: A Meta-Analysis. Experimental Economics 6, 299-310.

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

Fig. 1. Temporal structure of each period



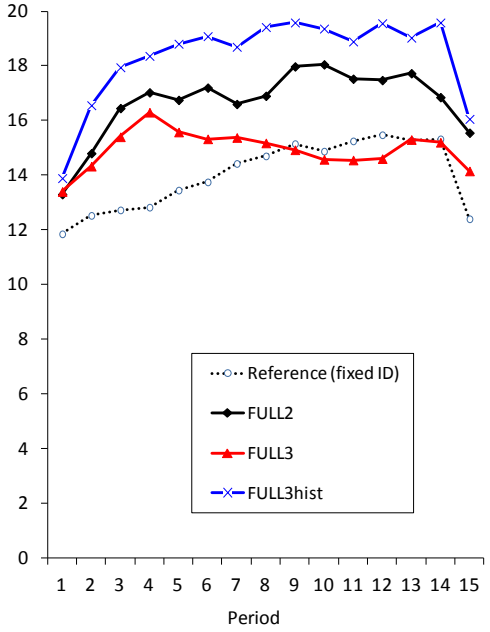
(a) Two Stage Periods (Reference, EGO2, FULL2)



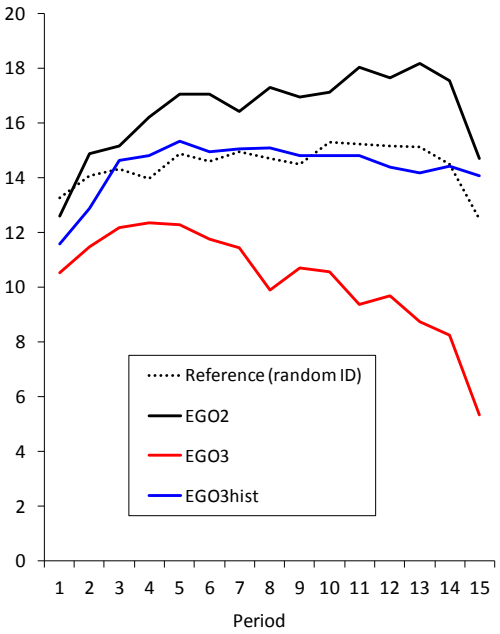
(b) Three Stage Periods (EGO3, EGO3hist, FULL3, FULL3hist)

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

(a) Average Contribution

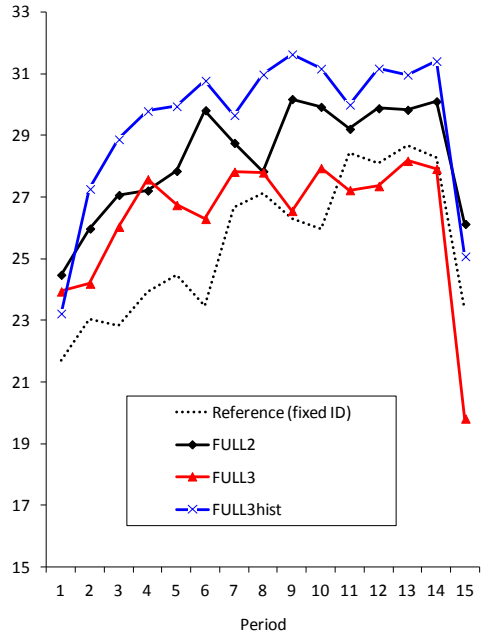


(i) Treatments with Full Information and Reference

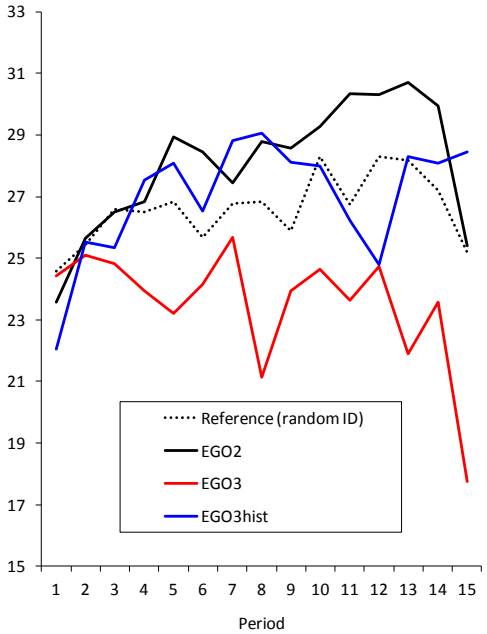


(ii) Treatments with Ego-centered Information and Reference

(b) Average Earnings



(i) Treatments with Full Information and Reference



(ii) Treatments with Ego-centered Information and Reference

Table 1. Summary of treatments, contributions and earnings

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[3]	Average contributions	Average Earnings
(a) Treatments with Full Information and Fixed ID Reference								
Reference (Fixed ID)	N	2	NO	NO	2	10	14.0	25.5
FULL2	F	2	YES	YES	2	10	16.7	28.3
FULL3	F	3	NO	YES	2	10	14.9	26.4
FULL3hist	F	3	YES	YES	2	9	18.3	29.5
(b) Treatments with Ego-centered Information and Random ID Reference								
Reference (Random ID)	N	2	NO	NO	2	10	14.5	26.6
EGO2	E	2	YES	YES	2	10	16.5	28.0
EGO3	E	3	NO	YES	2	10	10.3	23.5
EGO3hist	E	3	YES	YES	3	12	13.7	26.6
Experiment as a whole					17	81		

Notes: [1]N = no information on who punished whom, E = "Ego-centered information," F = "Full information."
[2] YES indicates that history of all past periods' contributions and punishments are displayed, except in treatment EGO2, where history information is available for the most recent period only. In EGO2 and EGO3hist, only history information concerning subjects who have punished the decision-maker is displayed.
[3] Each group has 4 subjects.

Table 2. Information and Punishment Opportunities 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*	Who Subjects are Permitted to Punish in Stage 3
Reference (Random ID)	No Information	Contribution decisions in period *t*	N.A.	N.A.
EGO2	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.	N.A.
EGO3	No Information	Contribution decisions in period *t*	Stage 2 punishment decisions of group members who have punished you in period *t*	Those who punished them in Stage 2
EGO3hist	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*	Those who punished them in Stage 2
Reference (Fixed ID)	No Information	Contribution decisions in period *t*	N.A.	N.A.
FULL2	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.	N.A.
FULL3	No Information	Contribution decisions in period *t*	Stage 2 punishment decisions of all group members in period *t*	Every individual in their groups
FULL3hist	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	Every individual in their groups

Table 3. Cost and distribution of punishment, by stage

	(a) Treatments with FULL Information & Ref.				(b) Treatments with EGO-centered Information & Ref.			
	Reference (Fixed ID)	FULL2	FULL3	FULL3hist	Reference (Random ID)	EGO2	EGO3	EGO3hist
(i) 2nd stage (1st order) pun. cost (per subject, per period)	2.91	1.72	0.74	1.03	2.09	1.84	2.17	1.34
share anti-social, as % of cost	*16.9%*	*15.8%*	*12.6%*	*11.5%*	*15.0%*	*20.6%*	*20.9%*	*10.3%*
share perverse, as % of cost	*17.8%*	*15.4%*	*13.5%*	*12.9%*	*28.8%*	*16.6%*	*15.7%*	*11.5%*
share anti-social, as % of events	*22.3%*	*19.7%*	*14.1%*	*16.5%*	*22.6%*	*23.6%*	*9.8%*	*8.3%*
share perverse, as % of events	*22.3%*	*21.1%*	*15.5%*	*18.8%*	*32.7%*	*20.1%*	*8.4%*	*10.0%*
(ii) 3rd stage (2nd order) pun. cost (per subject, per period)	----	----	1.87	0.50	----	----	0.51	0.23
relative strength of counter-punishment to anti-social vs. pro-social punishers	----	----	*4.16*	*15.38*	----	----	*0.63*	*0.92*
relative strength of counter-punishment to perverse vs. normal punishers	----	----	*3.08*	*6.72*	----	----	*0.54*	*1.28*
relative frequency of counter-punishing anti-social vs. pro-social punishers	----	----	*2.44*	*7.61*	----	----	*1.84*	*1.27*
relative frequency of counter-punishing perverse vs. normal punishers	----	----	*2.18*	*4.31*	----	----	*0.99*	*1.44*
Total punishment cost (per subject, per period)	2.91	1.72	2.61	1.53	2.09	1.84	2.68	1.57

Notes: Average total cost of punishment per period per subject is calculated as the cost to punisher plus cost to punishment recipient. Punishment of *i* by *j* is defined as anti-social if $C_i \geq C_j$ and pro-social if $C_i < C_j$. Punishment of *i* by *j* is defined as perverse if $C_i \geq$ the group's average contribution and as normal if $C_i <$ the group's average contribution. Relative strength of counter-punishment to anti-social vs. pro-social punishers is the ratio of the average number of counter-punishment points per point of anti-social punishment to the average number of counter-punishment points per point of pro-social punishment. Relative strength of counter-punishment to perverse vs. normal punishers is defined correspondingly. Relative frequency of counter-punishing anti-social vs. pro-social punishers is the ratio of the % of anti-social punishment events that are counter-punished to the % of pro-social punishment events counter-punished (or the ratio of the lengths of the relevant bars in Fig. 3(a)). Relative frequency of counter-punishing perverse vs. normal punishers is defined correspondingly (and can likewise be interpreted as the ratio of corresponding bar lengths in Fig. 3(b)).

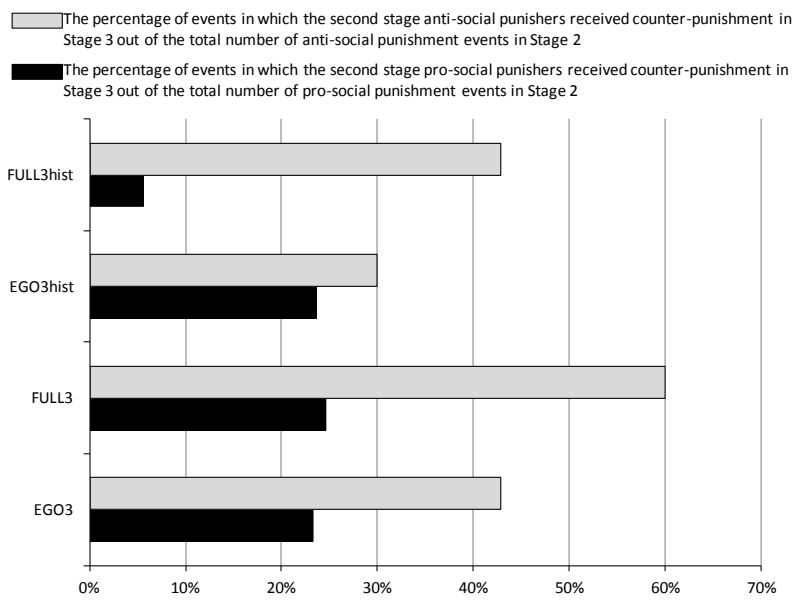
Table 4. Determinants of higher-order punishment received in EGO2 and FULL2 (Random Effects Tobit Regression)

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

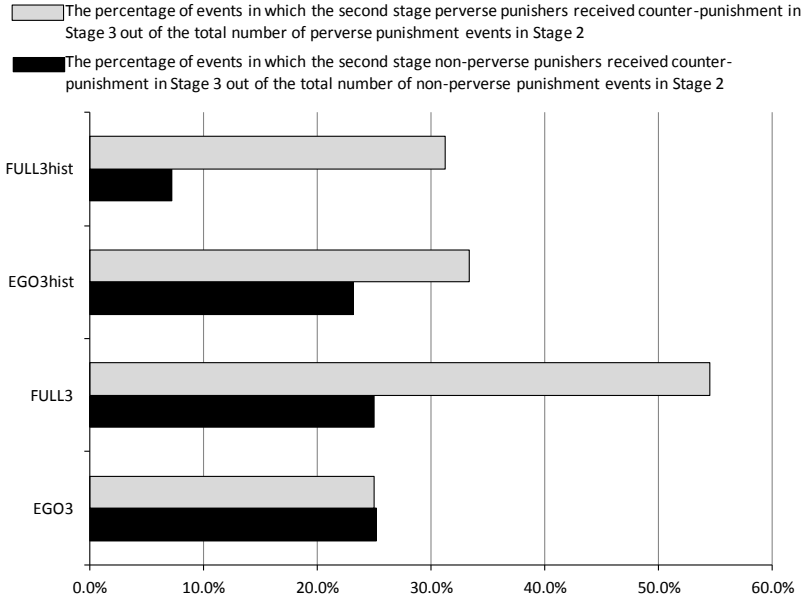
Independent Variable	EGO2		FULL2
	[1]	[2]	[3]
(i) Average contribution in group in period *t*	-0.16***	-0.083**	-0.23***
	(0.021)	(0.037)	(0.031)
(ii) Max {($C_{it} - C_{jt}$), 0} [abs. neg. deviation in *t*]	0.23***	0.058	0.26***
	(0.024)	(0.076)	(0.028)
(iii) Max {($C_{jt} - C_{it}$), 0} [pos. deviation in *t*]	0.033	0.014	0.057
	(0.030)	(0.055)	(0.036)
(iv) Max {($C_{i(t-1)} - C_{j(t-1)}$), 0} [abs. neg. deviation in *t* - 1]	----	-0.13	0.017
		(0.10)	(0.029)
(v) Max {($C_{j(t-1)} - C_{i(t-1)}$), 0} [pos. deviation in *t*- 1]	----	-0.036	-0.043
		(0.062)	(0.048)
(vi) $p_{ji(t-1)}$ if $C_{j(t-1)} > C_{i(t-1)}$, else 0	0.93***	0.43	0.75***
	(0.20)	(0.35)	(0.26)
(vii) $p_{ji(t-1)}$ if $C_{j(t-1)} \leq C_{i(t-1)}$, else 0	1.13***	1.47***	1.20***
	(0.28)	(0.41)	(0.46)
(viii) $\sum p_{jk(t-1)}$, all $k \neq i$ such that $C_{k(t-1)} \leq C_{i(t-1)}$	----	----	-0.49**
			(0.24)
(ix) $\sum p_{jk(t-1)}$, all $k \neq i$ such that $C_{k(t-1)} > C_{i(t-1)}$	----	----	0.00043
			(0.32)
(x) *t* [period]	-0.16***	-0.13**	-0.077**
	(0.029)	(0.068)	(0.033)
# of Observations	1680	162	1680
Log likelihood	-596.7	-127.0	-513.8
Wald Chi-squared	152.17	23.65	142.04
Prob > Wald Chi-squared	.000	.000	.000
Chi-squared Test on (iii) = (iv)			
Chi-squared	0.38	5.22	0.77
p-value	0.5402	.0223	0.3787

Notes: Random effects Tobit Regressions without constant term. 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. In column [2], only observations in which *j* punished *i* in the last period are used. The number of left- (right-) censored observations is 1528(1) in column [1], 125(0) in column [2], and 1549(0) in column [3]. *, **, and *** indicate significance at the .10 level, at the 0.05 level and at the .01 level, respectively.

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



(a) Pro-social versus Anti-social Punishment[#1]



(b) Non-Perverse versus Perverse Punishment[#2]

Notes: [#1] We call the punishment given by subject *i* "pro-social" if it is directed at those who contributed less than the contribution of subject *i*. By contrast, we call the punishment given by subject *i* "anti-social" if it is directed at those who contributed more than or equal to the contribution of subject *i*.

[#2] We call the punishment "non-perverse" if it is directed at those who contributed less than the punisher's contribution. By contrast, the punishment is "perverse" if it is directed at those who contributed more than or equal to the punisher's contribution.

Table 5. Determinants of higher-order punishment received in EGO3, EGO3hist, FULL3 and FULL3hist

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

Independent variable	EGO3 (1)	EGO3hist (2)	FULL3 (3)	FULL3hist (4)
(i) Average contribution in group in period *t*	0.038 (0.081)	-0.068 (0.0042)	-0.37*** (0.046)	-0.45*** (0.070)
(ii) Max {($C_{it} - C_{jt}$), 0} [abs. neg. deviation in *t*]	-0.099 (0.26)	-0.060 (0.084)	0.22*** (0.041)	0.13*** (0.040)
(iii) Max {($C_{jt} - C_{it}$), 0} [pos. deviation in *t*]	-0.15 (0.078)	-0.14 (0.053)	-0.078 (0.058)	-0.24** (0.12)
(vi) p_{jit} if $C_{jt} > C_{it}$, else 0	-0.010 (0.21)	0.23 (0.17)	2.01*** (0.46)	1.70*** (0.60)
(v) p_{jit} if $C_{jt} \leq C_{it}$, else 0	-0.11 (0.26)	0.10 (0.25)	1.07 (1.32)	3.61*** (0.87)
(vi) $\sum p_{jkt}$, all $k \neq i$ such that $C_{kt} \leq C_{it}$	----	----	-1.03** (0.48)	-0.47 (0.35)
(vii) $\sum p_{jkt}$, all $k \neq i$ such that $C_{kt} > C_{it}$	----	----	0.15 (0.79)	-0.91 (1.08)
(viii) Period	-0.11 (0.081)	0.35*** (0.050)	-0.035*** (0.042)	0.15*** (0.051)
# of Observations	143	120	1800	1620
Log likelihood	-133.53	-95.15	-639.12	-187.56
Wald Chi-squared	10.35	14.84	94.61	45.46
Prob > Chi-squared	.1106	.0215	.000	.000
Chi-squared Test on (vi) = (v)				
Chi-squared	0.17	0.20	0.50	4.07
p-value (2-sided)	.6833	0.6509	.4781	.0437**

Notes: Random effects Tobit Regressions without constant term. The right-censoring limit is not specified. 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. The numbers of left-censored(right-censored) observations are 107(0) in column (1), 91(0) in column (2), 1662(3) in columns (3), and 1578(0) in column (4).

*, ** and *** indicate significance at the 0.10 level, at the 0.05 level and at the .01 level, respectively.

Table 6. Determination of contributions and earnings by treatment settings

Linear regressions. Independent Variable	Dependent variable: Avg. contribution by subject *i*	Dependent variable: Avg. earnings of subject *i*
T: Third Stage Punishment Dummy (1 if EGO3, EGO3hist, FULL3, FULL3hist; 0 otherwise)	-4.17*** (1.07)	-3.09*** (0.82)
F: Full Information Dummy (1 if F, FULL3, FULL3hist; 0 otherwise)	0.67 (2.16)	1.44 (1.64)
H: History Dummy (1 if EGO2, FULL2, EGO3hist, FULL3hist)	1.99* (1.08)	1.45* (0.82)
T * F	4.44*** (1.51)	2.53** (1.15)
T * H	1.37 (1.49)	1.74 (1.14)
F * H	0.015 (1.51)	-0.086 (1.15)
Fixed ID dummy	-0.48 (1.08)	-1.11 (0.82)
Constant	14.48*** (0.76)	26.6*** (0.58)
# of Observations	324	324
F	9.85	9.74
Prob > F	.000	.000
Adjusted R-Squared	.161	.159

Notes: *, **, and *** indicate significance at the .10 level, at the 0.05 level and at the .01 level, respectively. No clustering of errors by group, for reasons discussed in the text. Results are considered suggestive, only.