In Broad Daylight: Fuller Information and Higher-Order Punishment Opportunities Can Promote Cooperation

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

[1] Department of Economics and Finance, Durham University, Mill Hill Lane, Durham, DH1 3LB, UK. Email: kenju.kamei@gmail.com, kenju.kamei@durham.ac.uk.

[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 this period, designated opportunities to counter-punish are harmful to cooperation and efficiency.
- Symmetric higher-order punishing opportunities with fuller information and history display aids cooperation significantly.

1. Introduction

In 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 greater cooperation levels (Herrmann *et al.*, 2008). An important unsettled question, however, is when (if ever) do the benefits of punishment opportunities survive the possibility of counter-punishment, feuds and vendettas. If these problems are serious and common, they could explain why modern societies favor the use of formal enforcement mechanisms and attempt to suppress peer-to-peer punishment (Markussen *et al*. 2014; Kamei *et al*., 2015). We contribute to the discussion by reporting a set of experiments that make available the information and opportunities with which to engage in higher-order punishment, including counter-punishment.

Views on the role of higher-order punishment cover a wide range, beginning with suggestions in the evolutionary literature that a preference for punishing non-cooperators (that is, for engaging in first-order punishment) could have been selected for thanks to second-order punishment of those who failed to (first-order) punish (Henrich and Boyd, 2001; Henrich, 2004).[1] On the other end of the spectrum are some 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. The concern about retaliation and vendettas was raised as early as John Locke's *Two Treatises of Government* (2005 [1689]) and used as one basis of his argument that sanctioning should be the province of government rather than of individual citizens. The potential of counter-punishment to deter and thus to undermine the efficacy of punishment, while adding to its cost, has been demonstrated in laboratory experiments by Denant-Boemont *et al*. (2007), Nikiforakis (2008), Hopfensitz and Reuben (2009), and Engel *et*

[1] Axelrod (1986) relatedly discusses "a norm that one must punish those who do not punish a defection," labeling it a "meta-norm."

al. (2011).[2] Nikiforakis and Engelmann (2011), Nicklisch and Wolff (2011), Nikiforakis, Noussair and Wilkening (2012), Engelmann and Nikiforakis (forthcoming) and Bolle, Tan and Zizzo (2010) further investigate how feuds can be resource-destroying and how the danger of feuds can reduce the frequency or increase the severity of first-order punishment 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 or that it will be used to retaliate against the punisher, 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 (second-order) punishment of (first-order) *non-punishers* as *punishment enforcement for omission* (PEO—i.e., for omitting to punish) and to (second-order) punishment of those who (first-order) punish cooperators 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 papers cited, our starting point is a multi-player, finitely repeated linear voluntary contribution mechanism (VCM, also known as public goods game), modified so that each period includes a post-contribution stage in which group members learn one another's contributions to the public good and have the chance to punish one another at some cost. In a reference treatment resembling "first generation" designs (e.g. Fehr and Gächter, 2000), group members are not

[2] Balafoutas and Nikiforakis (2012) and Balafoutas *et al*. (2014) also report that fear of counter-punishment is a deterrent to punishment in their field experiments on norm-violation and punishment. We note that their applications involve strangers in large urban settings, whereas some of the literature on punishment and cooperation among peers has in mind smaller group interactions, for instance irrigation associations, business partnerships, or small cooperatives. Our experiment was designed with such smaller group settings in mind.

informed of who punished them, and subject identifiers are scrambled each period to avoid vendettas. We conduct such treatments, then conduct additional treatments to study the effect of opportunities to engage in higher-order punishment.

Our additional treatments differ in four dimensions, described in the next section in more detail. These are (a) whether information and higher-order punishment opportunities are "ego-centric"—restricted to knowing who punished you and punishing them back—or cover the full set of potential dyads, allowing, e.g. an individual *k* to punish, say, *j* for punishing or failing to punish *i*; (b) whether group members are identifiable across periods by virtue of having fixed ID numbers; (c) whether there are or are not distinct decision stages in which higher-order punishment is the only available action; and (d) whether subjects are provided with explicit reminders of one another's past punishment actions when making punishment decisions.

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. In one treatment, which provides full information, a dedicated higher-order punishment stage, generalized higher-order punishment opportunities, and explicit display of information on past behaviors, contributions and earnings are significantly higher than in the corresponding reference treatment. The sole treatment in which information and additional punishment opportunities prove harmful, in contrast, 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. We consider broader implications for the efficacy of peer punishment in the face of higher-order-punishment opportunities, in our concluding discussion.

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 and concludes.

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 VCM, each period begins with subjects receiving an endowment of experimental currency and independently making first-stage decisions on what, if anything, to contribute to a group account. The period then has 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. A subject *i* has earnings in period *t* 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 *i*'s allocation to the public good in period t, β is the effectiveness of punishment, 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 $j \neq i$ gives to subject i in period t. The term in brackets is subject *i*'s earnings from the allocation stage, the second term her loss from being punished, and the third term 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 t whereas maximization of own payoff entails $C_{it} = 0$ for all i and t. In one-shot play or in the last period 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 (by backward induction) if there is common knowledge that all group members are rational maximizers of own payoff.

In finitely repeated laboratory public goods experiments *without* punishment opportunities, the prediction that $C_{it} = 0$ fails, with contribution averaging between 40 and 60% of endowment in the initial period, then declining more or less monotonically with repetition (Zelmer, 2003). When costly punishment is available, enough subjects pay for punishment that is consistently enough targeted at lower contributors, at least in populations initially studied, that contributions decline more slowly than in punishment-free settings, or even rise with repetition.[3]

In first generation VCM experiments with punishment, subjects are not informed who punished them by what amount, and identifiers are switched from period to period to discourage vendettas. Subjects also lack information about others' punishing practices, which along with the identification changes means that punishment enforcement for omitting to punish (PEO) and for committing wrongful (perverse) punishment (PEC) are ruled out. Some subjects, however, appear to 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 have punished her (Cinyabuguma *et al*., 2006; Herrmann *et al*., 2008). One might thus anticipate that providing the information on which counter-punishment can be based would lead to more such revenge, lowering efficiency. The results obtained in the first treatments permitting counter-punishment support this (Denant-Boemont *et al*., 2007; Nikiforakis, 2008).

We designed new treatments of finitely repeated voluntary contribution experiments motivated by the concern that the apparent power of peer-to-peer punishment to stabilize cooperation in first-generation cooperation and punishment experiments might be misleading because they artificially shield subjects from punishment conditioned on punishing, as opposed to contributing, behaviors. More than one new treatment is required because there is room for debate exactly how the shielding should be removed. 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 without subsequent

[3] This rising trend is more likely the higher is β (Nikiforakis and Normann, 2008). Experiments reported in Herrmann *et al*. 2008 suggest that early results, obtained mainly using subjects in western European and U.S. universities, do not predict well outcomes in some southeast and east European and Middle Eastern subject pools.

display of the required information.[4] We vary history and identifiability along the spectrum from conditions lacking both identifiability of punishers and future display of history 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 subsequent display of information to aid memory.

Information and punishment restrictions. Some experimental treatments including the Counter-punishment treatment of Nikiforakis (2008) and the Revenge Only treatment of Denant-Boemont *et al*. (2007) restrict information on punishment and opportunities for higher-order punishment in a manner we describe as "ego-centric," meaning that a subject learns only of those punishments directed at herself, and can punish the punishing acts only of her own punishers. These designs rule out the enforcement acts by other group members that we've called PEO and PEC. A justification for this may be that, especially in larger groups, one may be able to readily observe punishments to and from oneself only. But conditions of observability are situation-specific, making it reasonable to consider the ego-centric restrictions (only *j* observes p_{ij}) as part of a continuum of possibilities. To one side of it lies the situation of complete non-observability of who the punisher is, as in first generation cooperation and punishment experiments. At the opposite end of the continuum lies equal observability of punishment interactions between any two group members (*i* and *j*, *j* and *k*, etc.), a condition studied experimentally in Denant-Beomont *et al.*'s Full Information treatment and in Nikiforakis and Engelmann (2011) and other papers on feuding (see below). Our experimental treatments include ones representing each of these three configurations.[5]

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

[4] We refer to subsequent information display as distinct from ID maintenance because we include treatments in which IDs are scrambled, but subjects are nonetheless presented with selected information about other group members' past behaviors; see the discussions of treatments EGO2 and EGO3Hist, below. Such information display makes it possible to identify past punishments one received from specific individuals, yet leaves other information (such as the past contributions of those group members who have not been one's punisher) opaque.

[5] While our range of treatments resembles that in Denant-Beomont *et al*. in this respect, there are differences in some other design features and the important difference of outcome that unlike their treatments, several of ours suggest that permitting higher-order punishment aids cooperation, with this advantage being statistically significant for one treatment.

counter-punishment stage introduced in Nikiforakis (2008). In real world interactions, punishment chains reaching up to very high orders may take place. One way to accommodate this in the lab is to allow other activity to stop until subjects have engaged in as many rounds of punishment and counter-punishment as they wish to so long as resource constraints are obeyed, as in Nikiforakis and Engelmann (2011) and other studies of feuding. However, suppose that one's reference point is something like 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 as work continues. Here, the pausing of first stage decisions for multiple rounds of punishment may be impracticable. While an individual might choose to continue an ongoing feud with another group member for an extended period of time, the fact that the first-order decisions (working or contributing, or not doing so) which occasioned the punishment in the first place continue at intervals within that extended time may alter outcomes. One reason is that if higher-order punishment follows intervening contribution decisions, observers including the punished individual can't cleanly decompose its causes. Another is that returning to the group's first-order business before resuming punishment may cool the urge to punish, either because it gives the prospective target an opportunity to show reformed behavior, or simply due to fading of attention or emotion.

Because the approach of pausing first-order collective action for multiple punishment stages has been explored extensively by others and because situations that it may fail to mirror well strike us as common and important, we study higher-order punishment in a repeated play dilemma game with the constraint that the number of punishment stages not exceed the number of contribution stages by more than a factor of two. Specifically, we consider two kinds of set-up: (1) ones in which each period has a dedicated higher-order punishment stage following first-order punishment, thus a total of three stages (contribution, punishment, punishment of punishment), and (2) ones in which each period offers only one punishment stage, so higher-order punishment is possible but only by taking the punishments of past periods into account after the contribution decisions of later periods. With sufficient continuity of identity and reminders of past actions, neither approach strikes us as too strongly protecting subjects against punishment of higher order, and we think our constraint useful to understanding an important subset of situations in which individuals may try to cooperate toward some end without benefit of external authority.

2.2 The experiment

Our experiment begins with a Reference (random ID) treatment of the standard first-generation cooperation and punishment variety, with a single punishment opportunity and scrambling of identifiers to prevent higher-order punishment apart from "blind revenge." Along with it, we conduct six treatments providing information-based higher-order punishment opportunities. Because identifiers are not scrambled in three of those six treatments (the FULL treatments, see below), we add a Reference (fixed ID) treatment that has no higher-order punishment opportunities but in which identifiers remain fixed 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 knowingly without change of partners and for a total of fifteen periods. As in many symmetric, linear VCM set-ups, each subject gets a new endowment to allocate each period and earns one point for each unit he places in his private account, while generating 0.4 for himself and for each other group member with each point allocated to the public good, a value facilitating comparability with Fehr and Gachter (2000, 2002), Nikiforakis (2008) and many others among the papers mentioned above. Thus, subjects can earn up to 1.6 times as much if all fully contribute than if they each allocate nothing in the group account. To simplify instructions and interpretation, we use a fixed ratio of punishee loss to punisher cost (parameter β of Eq. (1) above): a punisher pays one point to reduce the earnings of a targeted group member by three points ($\beta = 3$). 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 the traditional assumption of common knowledge of (own) payoff maximizing type, subjects always incur the cost of any punishing they themselves chose to impose.[6]

In the four treatments in which each period has a dedicated higher-order punishment stage—our 3 stage treatments—earnings of a subject *i* in period *t* are given by:

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

where p_{jit} is the amount of punishment *i* receives from $j \neq i$ in the first punishment stage, pp_{jit} is third stage (2nd-order) punishment of *i* by subjects $j \neq i$, S_{jt} is the set of group members $j \neq i$ that are permitted to punish *i* in the third stage, and S_{it} is the set of group members whom *i* can higher-order punish. The term in large brackets is thus *i*'s earnings net of punishment received, with floor at 0, while the remaining two terms represent *i*'s expenditures (if any) on first and second opportunity (also called 2nd and 3rd stage) punishing.

In two of the 3 stage treatments, dubbed EGO3 and EGO3hist, S_{jt} contains only group members that *i* punished in stage 2 and S_{it} likewise contains only group members who engaged in stage 2 punishment of *i*, presumably punishing *i* for her contribution decision. In the others, dubbed FULL3 and FULL3hist, S_{jt} and S_{it} include all group members $j \neq i$. In EGO treatments, subjects' IDs and screen positions are scrambled each period, whereas in FULL treatments, they remain fixed throughout the 15 periods of interaction. Indeed, in all treatments with higher-order punishment opportunities including the two-stage treatments described below, FULL and EGO treatments differ from each other on three counts: (i) whether identities remain fixed (FULL) or not (EGO), (ii) whose punishment one is informed of (all, in FULL, that directed at oneself only, in EGO), and (iii) who one can give (and receive) higher-order punishment to (from)—any

[6] Net losses, in practice rare and limited to a few periods, are covered out of earnings from other periods. 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.

group member in FULL, those who punished one only, in EGO.[7] The "hist" suffix indicates display of averaged past behaviors, and is detailed further, below.

A possible concern about having 3 stages in a period is that adding a stage in which a subject's only available action is counter- or (other) higher order-punishing may invite more such punishing due to an "experimenter demand effect," relative salience, or "hot anger."[8] To deal with these possibilities, we add two treatments in which periods have only one contribution and one punishment stage—2 stage treatments. Higher order punishment is possible in the second and later periods of these treatments since subjects decide on punishment while seeing information on both current period contribution and previous periods' punishments. But there is only one punishment act per period, so its clean decomposition into first- and higher-order components is impossible both for subjects and observers. In EGO2, subjects are shown at the punishment stage of periods $t \geq 2$ each $p_{ji(t-1)} > 0$—the same information available to counterparts in the third stage of periods in EGO3 and in Nikiforakis (2008)'s counter-punishment treatment. They are also shown their punisher's period t contribution and reminded of their own, which might put the punishment in perspective. In the punishment stage of the same periods in FULL2, subjects see a display not only of all bilateral punishments in their group in period $t - 1$, but also the average past punishment between each pair in the group from period 1 to $t - 2$.[9]

Unfortunately, when comparing EGO3 to EGO2 and FULL3 to FULL2, it is difficult to distinguish the impact of displaying past period actions, something required for 2 stage treatments with counter-punishment, from the effect of having 2 vs. 3 stages. This concern led us to add the "hist" treatments mentioned above. Those treatments are distinguished from EGO3 and FULL3 by display of history only. In FULL3hist, subjects see information at the

[7] In principle, the three dimensions could be separated; e.g., subjects could be shown the full patterns of 1st order punishments but then permitted to counter-punish their own punishers only in a 3rd stage. Because our experiment is already quite complex and this combination of features seemed natural to us as a starting point, we leave it to future research to investigate such further unpacking of design dimensions.

[8] Subjects are free to choose 0 for each pp_{ij}, but the concern about experimenter demand is that some may view that choice as passive, and that since they came to the lab to be paid for making decisions, they might lean towards seemingly more active ones. See Zizzo (2009). To be sure, the fact that setting $pp_{ij} > 0$ is costly should act as a countervailing factor.

[9] The earlier average information is available from period 3 onwards. We keep depth of history shallow in EGO2 to permit us to isolate any experimenter demand, hot anger, or other effects that might be attributable to differences with EGO3.

contribution and punishment stages of later periods regarding both all bilateral punishments in the previous period (t – 1) and the average of such punishments in periods prior to that (periods 1 to t – 2). This display can in principle inform FULL3hist subjects of who punished them and by how much in the previous period's third stage, information unavailable to FULL3 subjects, and the display of history in later periods might also make a difference by aiding over-taxed memories or raising salience. In EGO3hist, which has a third stage for punishing based on those first-order punishments (of oneself) that are reported to a subject, there is similar display of information about punishments in the most recent and average punishments in all earlier periods, but with the "ego-centric" property that only those past punishments aimed at oneself are displayed. A schematic depiction of the structure of 2 Stage and 3 Stage periods is provided in Fig. 1, while Table 2 provides details about the information available to subjects at the various stages of a period, by treatment. Full experiment instructions, including examples of the informational screens viewed by subjects in each treatment, are included in online Appendix C.

3. Results

17 experiment sessions were conducted in a computer lab at Brown University between October, 2011 and January, 2013.[10] Table 1 shows numbers of sessions and groups by treatment. In all, 324 student subjects participated.[11] Sessions typically took 75 to 90 minutes from signing of consent forms to reading instructions aloud and (simultaneously) from printed copies, answering 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 and Reference (fixed ID) treatments in the left panels and EGO and Reference (random ID) treatments in the right panels. Both Reference treatments display the

[10] 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, dropping the data of the compromised groups.

[11] Subjects were recruited from the general undergraduate population of Brown University, representing majors in the humanities, social sciences, and sciences, with 19.6% being economics majors (slightly higher than their share in the general student population) and 52.9% female (also slightly above a representative share).

familiar contributions pattern of first-generation contribution and punishment experiments. The average contribution begins around 60% of endowment, and then trends upwards towards 75% of endowment before a last-period decline.[12]

In the left panels of Fig. 2, the three treatments permitting informed higher order punishment all begin with and continue to display higher average contributions and earnings than their reference counterpart for at least the first seven periods, with average contributions and earnings remaining higher than in Reference (fixed ID) for all periods of FULL3hist and FULL2. Mann-Whitney tests using group-level observations of average contribution for periods 1 – 15 as a whole find contributions and earnings to be statistically significantly different from Reference (fixed ID) for the treatment having the highest average contribution curve, FULL3hist ($p = .014 < 0.05$, 2-tailed test).[13]

Corresponding comparisons of the EGO treatments in Fig. 2's right panels suggests that one treatment, EGO2, has somewhat higher contributions and earnings, another, EGO3, considerably lower contributions and earnings, and the third, EGO3hist, roughly similar contributions and earnings relative to the benchmark treatment, Reference (random ID). Average contributions and average earnings for the 15 periods as a whole are statistically significantly lower in EGO3 than in EGO2 and Reference (random ID).[14] The performance difference of EGO2 versus Reference (random ID) is not statistically significant, overall. While comparisons of fixed and random ID treatments are potentially problematic, hence otherwise avoided in our discussion, we mention for completeness that contributions and earnings are statistically significantly lower in EGO3 than in all three FULL treatments.

[12] Insofar as Fehr and Gächter (2000) scrambled identifiers to prevent reputation formation and 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, there are no statistically significant differences between the two treatments with respect to total punishing events or punishment, total punishment given by lower to higher contributors (perverse punishment), or share of the opportunities in which a lower contributor could punish a higher one that were actually utilized. There are also no significant differences in contribution amount or earnings.

[13] Contributions in FULL3hist are also significantly different than those in FULL3 ($p = 0.034 < 0.05$, 2-tailed test). Contributions and earnings overall are not statistically significantly different between other treatments. See Appendix Table B.2 for details.

[14] $p = 0.005$ and 0.082, respectively, 2-tailed tests for contributions, and $p = 0.003$, and 0.082, respectively, 2-tailed tests of earnings. Earnings are also significantly lower in EGO3 than in EGO3hist: $p = 0.048$, 2-tailed; see again Appendix Table B.2.

Result 1: *Allowing higher-order punishment significantly raises contributions and earnings in a treatment with full information, a dedicated higher-order punishment stage, and display of past history, and significantly lowers contributions and earnings in a treatment with ego-centric information, a dedicated higher-order punishment stage, and no display of past history. Average contributions and earnings are not significantly changed by adding the possibility of higher-order punishment to that of first-order punishment in our other four treatments.*

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 (PEO), and sanction enforcement for punishing cooperators (PEC). Looking first at the cost of punishing, measured as the sum of costs incurred by punishers and by those targeted for punishment, Table 3 shows considerable variation among treatments with regard to amounts lost to stage 2 (first-order) punishing: almost four times as much in Reference (fixed ID) as in FULL3, in which it appears that subjects often delayed punishment until stage 3 so as to reduce the likelihood of retaliation.[15] The bottom row of Table 3 shows that comparing total punishment per period, adding together both stages in the 3 stage periods, narrows overall differences to a gap of just 11% in the two treatments just mentioned and to a maximum gap of 1.9:1 between the Reference (fixed ID) and the two "hist" treatments. Impressionistically, availability of higher-order punishment opportunities seems to discourage first-order punishing, with the difference made up by second-order (stage 3) punishment in FULL3 and EGO3 but less so in their "hist" counterparts. 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 total punishment (see Appendix Table B.3).[16]

[15] Escaping retaliation is possible in FULL3 because subjects are not shown who gave what stage 3 punishment in that treatment. While subjects could delay first-order punishing in FULL3, this was not possible in EGO3 because there a subject can punish in the third stage only a group member who has punished her in the same period's second stage.

[16] 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. Regarding share of Stage 2 punishment opportunities utilized, 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

With total amounts of punishing not dramatically different, understanding what drives differences in outcomes by treatment requires a closer look at punishing patterns. The italicized rows in the upper portion of Table 3 give information about the pattern of stage 2 (mainly first-order) punishment, while corresponding rows in the lower portion (under heading (ii)) give information about the pattern within that part of stage 3 (mainly second-order) punishment that can be classified as counter-punishing because *j* is punishing *i* immediately following *i* punishing *j*. Overall, 82.6% (by cost) of stage 2 punishment was targeted at low contributors, and the proportions of first-order punishment expenditures and events that are in some sense directed at cooperators, and more precisely speaking are classifiable as anti-social or perverse,[17] fail to show a clear pattern of differentiation between EGO and FULL treatments. For example, the two shares are somewhat higher in the EGO3 than in the FULL3 treatment, but they are slightly higher in FULL3hist than in EGO3hist, and shares of anti-social and perverse punishment *events* (the third and fourth italicized rows) are greater both in FULL3 vs. EGO3 and in FULL3hist vs. EGO3hist.

The rows in part (ii) of Table 3, however, show interesting 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 on average when a perverse or anti-social punishment was being counter-punished than when a non-perverse or pro-social punishment was 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 non-perverse punishment in EGO3 and EGO3hist. Similar patterns apply in the bottom two italicized rows, which show ratios of (a) the *proportion* of anti-social or perverse punishment events counter-punished to (b) the corresponding proportion of pro-social or non-perverse

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.

[17] Following Herrmann *et al*. (2008) and Cinyabuguma *et al*. (2006), respectively, a punishment event is classified as anti-social when the punishing subject contributed less than the punished one, and as perverse when the punished subject contributed more than the average within the group for the period in question.

punishment events counter-punished. These relatively stronger responses to anti-social and perverse punishment in FULL than in EGO treatments are displayed graphically in Figure 3.[18]

To better understand the determinants of higher-order punishment, we estimate regression equations in which potential proximate causes for punishing appear as 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, so our regression analysis serves as an approximate way of teasing these components apart. Table 4 shows estimates of random effects Tobit regressions in which the observations are specific to each pair of subjects in a group.[19] The dependent variable is punishment received by subject j from subject i in period t, and the list of explanatory variables includes both current period factors likely to induce first-order punishing and past period factors that might induce second-order punishing. As in past studies, the first set of factors 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), all measured for period t (see variables (i) – (iii)) as potential reasons for first-order punishment. As the main factors expected to determine second-order punishment, the regressions include the amount of punishment j gave i in period $t - 1$ "pro-socially" (i.e., if $C_{j,t\text{-}1} > C_{i,t\text{-}1}$, variable (vi)) and the amount (if any) of "anti-social" punishment of i by j in that period (variable (vii)), these being distinguished to allow for differences in response. Since subjects in the FULL2 treatment were shown amounts of punishment given last period to group members other than themselves, regression [2], for FULL2, also controls for punishment j gave in period $t - 1$ to group members $k \neq i$ who contributed less than or the same amount as i (variable (viii)) and to any $k \neq i$ who contributed more than i (variable (ix)). Both regressions include a control

[18] The ratios of gray to black bar lengths for each treatments in this figure reflects the relative frequency of counter-punishing anti-social vs. pro-social punishers (panel (a)) and the relative frequency of counter-punishing perverse vs. normal punishers (panel (b)) in the bottom two italicized rows of Table 3. Unfortunately, non-parametric tests of whether the observed differences are statistically significant yield insignificant test statistics because only a few observations are fully defined. For details, see the note under Table B.3 of the Appendix.

[19] 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.

for period.[20] Finally, the deviations of *j*'s from *i*'s contribution in period $t - 1$ could potentially influence *i*'s decision to punish in *t*, so lagged deviation terms are added in the regression for FULL2, in which subjects have this information.[21]

The regressions for both treatments support the usual pattern of first-order punishment being significantly larger the greater the negative deviation between punisher and punishment target, while the positive deviation terms also obtain smaller and less significant positive coefficients. Presence of counter-punishment for last period's punishing is strongly supported by all four 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)), echoing findings in Table 3 and Figure 3. Punishment enforcement variable (viii) obtains a marginally significant negative coefficient, suggesting relative approval of pro-social punishment acts by subjects in FULL2.

For the 3 stage treatments, we use the regressions of Table 5 to study how stage 3 punishment responds to stage 2 punishment and other factors. 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, so only the relevant subset of observations are included, considerably restricting the sample.[22] 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

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

[21] Those EGO2 subjects *i* who were punished last period by the *j* in question also know the relevant period $t - 1$ deviation, but this applies to slightly under 10% of observations in a given period. Estimating a version of Table 4's regression [1] that includes only those observations and adds the two lagged deviation terms (i.e., variables (iv) and (v)) shows their coefficients to be insignificant.

[22] To avoid over-complicating the analysis, the analysis assumes selection bias (if any) to be of secondary importance.

and *j*'s contributions in that period are included.[23] As with Table 4's regression [2], specifications for the FULL treatments allow higher-order punishment to be conditioned also on any first-order punishment *j* gave or failed to give to third parties (PEC and PEO).

While the table shows only two coefficients to be statistically significant (and one of these only marginally so) in the regressions for the EGO treatments, there are many significant coefficients in those for the FULL treatments, a difference perhaps in part 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 significant positive coefficients on both the amount of pro-social and the amount of anti-social first-order punishment received by *i* (variables (iv) and (v) respectively). For FULL3hist, the coefficient for counter-punishing pro-social punishers is somewhat less than half as large as that for counter-punishing anti-social ones, consistent with previous remarks and with the impression conveyed by Figure 3. The absence of a significant coefficient for counter-punishment of anti-social punishers (variable (v)) in the FULL3 regression, in contrast, suggests a weakness of efficiency-promoting counter-punishing in that treatment, somewhat contrary to the impressions given by Table 3 and Figure 3. The clearer pattern of differential counter-punishment in the FULL3hist than in the FULL3 treatment is consistent with if not an independent cause of the higher contributions and efficiency achieved by subjects in FULL3hist.

The coefficients on sanction enforcement variables (vi) and (vii) are insignificant in Table 5, so the regressions provide no evidence of either PEO or PEC. We also used non-regression techniques to check the data for more direct signs of these kinds of sanction enforcement. In FULL3 and FULL3hist, we found cases in which a first-order (stage 2) non-punisher received punishment in stage 3, consistent with PEO. But the large majority of these cases can be explained as delayed first-order punishment. Observable PEC also turns out to be

[23] Whereas second-order punishment in period *t* of our two-stage treatments must of necessity reference the previous period, in the three-stage treatments it can occur in stage 3 with reference to the stage 2 punishments of the same period, so actions in period $t-1$ are likely to be less salient. To avoid clutter, we thus decided to omit lagged terms in Table 5's regressions.

rare, paralleling the findings reported by Denant-Boemont *et al.* (2007) and Nikiforakis and Engelmann (2011). Our data thus suggest to us that rather than widespread use of higher-order punishment opportunities for the purposes proposed by Henrich and Boyd (PEO) or for those suggested by Cinyabuguma *et al*. and Denant-Boemont *et al*. (PEC), the differences in induced cooperation and efficiency among our treatments are due mainly to the different patterns of counter-punishment by punishment recipients themselves. 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.

Result 2: *(a) We find evidence of counter-punishment by both direct and regression methods in both the 3-stage and the 2-stage treatments that permit informed higher-order punishment (i.e., those other than the Reference treatments). (b) Most of this evidence suggests that first-order punishment is more strongly counter-punished when perverse or anti-social in "FULL" treatments, where subjects are shown the full pattern of bi-lateral punishments, and is less differentially or not differentially counter-punished when perverse or anti-social in "EGO" treatments, where subjects are shown punishment given to themselves only, can only counter-punish their own punishers, and lack IDs that remain fixed across periods.*[24] *(c) We find little evidence of sanction enforcement, that is, of higher-order punishment by other group members predicated on punishments not directed at themselves.*

Subjects possess the requisite information and the opportunity to counter-punish in both FULL and EGO treatments, so what accounts for the difference in the relative strengths of pro- versus anti-social counter-punishing, and for the other differences in outcome, in these sets of treatments? 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

[24] The exception is the regression evidence for treatment FULL3.

complete information about the overall pattern of punishing in the group in FULL treatments, especially FULL3hist, 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, as well as better conveying group members' opinions about free riding in the contribution stage, which could have direct effects of its own on contribution choices. 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).

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 compared to EGO3, and contributions are also relatively high in FULL2. The presence of information on subjects' past play seems to raise efficiency in FULL3hist above that in FULL3, perhaps in part by reducing the incentive to delay punishment until stage 3, 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.[25]

Finally, to explain the sustaining of contributions and earnings in most of our higher-order punishment treatments, including their statistically significant enhancement in the FULL3hist treatment, versus the lack of increased contributions or efficiency given comparably symmetric information and higher-order punishment opportunities in the short and long feuds treatments 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

[25] We report in the Appendix an attempt to identify the effects of each of the three varying treatment dimensions by using ordinary least squares regressions on individual level observations. Having three stages obtains a significant negative coefficient, its interaction with FULL an offsetting significant positive coefficient, and displaying history a marginally significant positive coefficient. We view the results as suggestive only due to problems with the independence of observations; see Table B.12 for details.

reduces the duration of any feuding and reduces the concern over possible feuds as a disincentive to engaging in first-order punishment.[26]

4. Conclusions

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 the Counter-punishment treatment of Nikiforakis (2008) and the Revenge Only treatment of Denant-Boemont *et al*. (2007), we reconfirm that when subjects are shown information only about the amount of punishment they themselves receive from identifiable others, have a dedicated opportunity to punish back, and when identifiers last one period only, 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 after the next period's contribution stage), or provision of more depth of historical information, even though still ego-centric, prove sufficient in our settings to eliminate the negative effect of opportunities to engage in counter-punishment. When subjects in our experiment are in addition provided with fixed identifiers and with more general higher-order punishment opportunities, including but not limited to counter-punishment, efficiency in the presence of higher-order punishment opportunities tends to exceed, though not always significantly, that in a treatment with no such opportunities. 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, representing the maximum number of punishment stages before a new round of contributions, in our design.

Analysis of punishment patterns in those treatments permitting it suggests that higher-order punishment by third parties is rare and 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

[26] 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.

history retention. 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. We conjecture 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.

More broadly, our results suggest to us 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 available. Although peer-to-peer punishment may in many cases open the door to counter-punishments, there is also likely to be some observability of the pattern of punishment overall, and (in some subject pools, at least) norms can emerge wherein most group members understand that punishment of free-riders is applauded whereas punishment of cooperators is frowned upon. Full information regarding who punished whom combined with symmetric opportunities to engage in higher-order punishment and ongoing identifiability of individual group members aids the cooperation-enhancing effects of informal sanctions in our experiment.

Of course, small-scale laboratory experiments such as ours should be interpreted with great caution, their relevance to real world situations including collective action dilemmas played out on much larger scales being at most only suggestive. A more specific caveat is that unlike experiments on feuding, subjects in our experiment are not permitted to pause decision-making with respect to the (first-stage) collective action problem itself in order to engage in multiple retaliation rounds. Indeed, the applicability of our sanguine findings about the impact of higher-order punishment versus the more cautionary findings of feuding studies (e.g., Nikiforakis and Engelmann, 2011) may depend on whether the problem in question is characterized by ongoing collective action demands precluding prolonged strings of punishment and counter-punishment, or by ease of delaying collective action demands until feuds are played out, an environment that in the studies in question can give rise to heavy pre-emptive first punishment strikes.

Acknowledgments: We thank Jacob Murray, Iñaki Arbeloa and Yunan Ji for their help preparing and conducting the experiments. The Department of Economics at Brown University provided funding. Pilot treatments in collaboration with Jean-Robert Tyran (see Appendix C of Kamei *et al*., 2015) helped us launch this research. We thank Dirk Engelmann and two anonymous referees for helpful comments, and Hui Wen Ng and Xinyi Zhang for their assistance in revising earlier versions of the paper.

References

Axelrod, R., 1986. An evolutionary approach to norms. American Political Science Review 80, 1095-1111.

Balafoutas, L., Nikiforakis, N., 2012. Norm enforcement in the city: a natural field experiment. European Economic Review 56, 1773-1785.

Balafoutas, L., Nikiforakis, N., Rockenbach, B., 2014. Direct and indirect punishment among strangers in the field. Proceedings of the National Academy of Sciences, 111, 15924-15927.

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

Bowles, S., 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.

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.

Engelmann, D., Nikiforakis, N., forthcoming. In the long run we are all dead: on the benefits of peer punishment in rich environments. Social Choice and Welfare (in press).

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90, 980-994.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137-140.

Henrich, J., 2004. Cultural group selection, co-evolutionary 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., Reuben, E., 2009. The importance of emotions for the effectiveness of social punishment. Economic Journal 119, 1534-59.

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

Locke, J., 2005. [1689]. Two treatises of government and a letter concerning toleration. Digireads.com Publishing, Stilwell.

Markussen, T., Putterman, L., Tyran J.-R., 2014. Self-organization for collective action: an experimental study of voting on sanction regimes. Review of Economic Studies 81, 301–324.

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.

Nikiforakis, N., Engelmann, D., 2011. Altruistic punishment and the threat of feuds. Journal of Economic Behavior and Organization 78, 319–332.

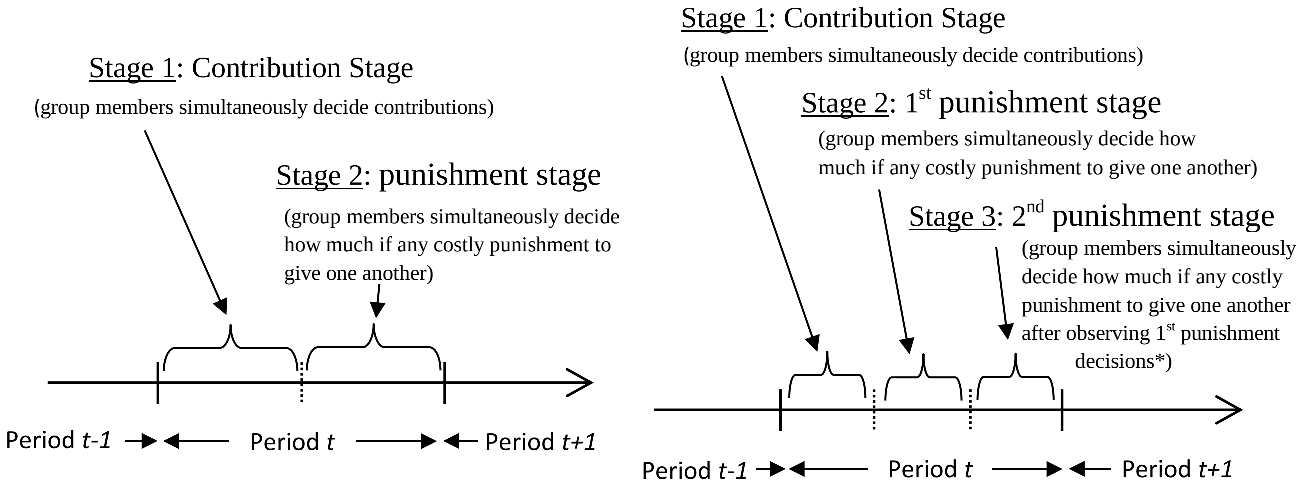
Nikiforakis, N., Normann, H.-T., 2008. A comparative statics analysis of punishment in public goods experiments. Experimental Economics 11, 358-369.

Nikiforakis, N., Noussair, C., Wilkening, T., 2012. Normative conflict and feuds: the limits of self-enforcement. Journal of Public Economics 96, 797–807.

Zelmer, J., 2003. Linear public goods experiments: a meta-analysis. Experimental Economics 6, 299-310.

Zizzo, D., 2009. Experimenter demand effects in economic experiments. Experimental Economics 13, 75-98.

Fig. 1. Temporal structure of each period


Stage 1: Contribution Stage
(group members simultaneously decide contributions)
Stage 2: punishment stage
(group members simultaneously decide how much if any costly punishment to give one another)
Period t-1
Period t
Period t+1
Stage 1: Contribution Stage
(group members simultaneously decide contributions)
Stage 2: 1st punishment stage
(group members simultaneously decide how much if any costly punishment to give one another)
Stage 3: 2nd punishment stage
(group members simultaneously decide how much if any costly punishment to give one another after observing 1st punishment decisions*)
Period t-1
Period t
Period t+1

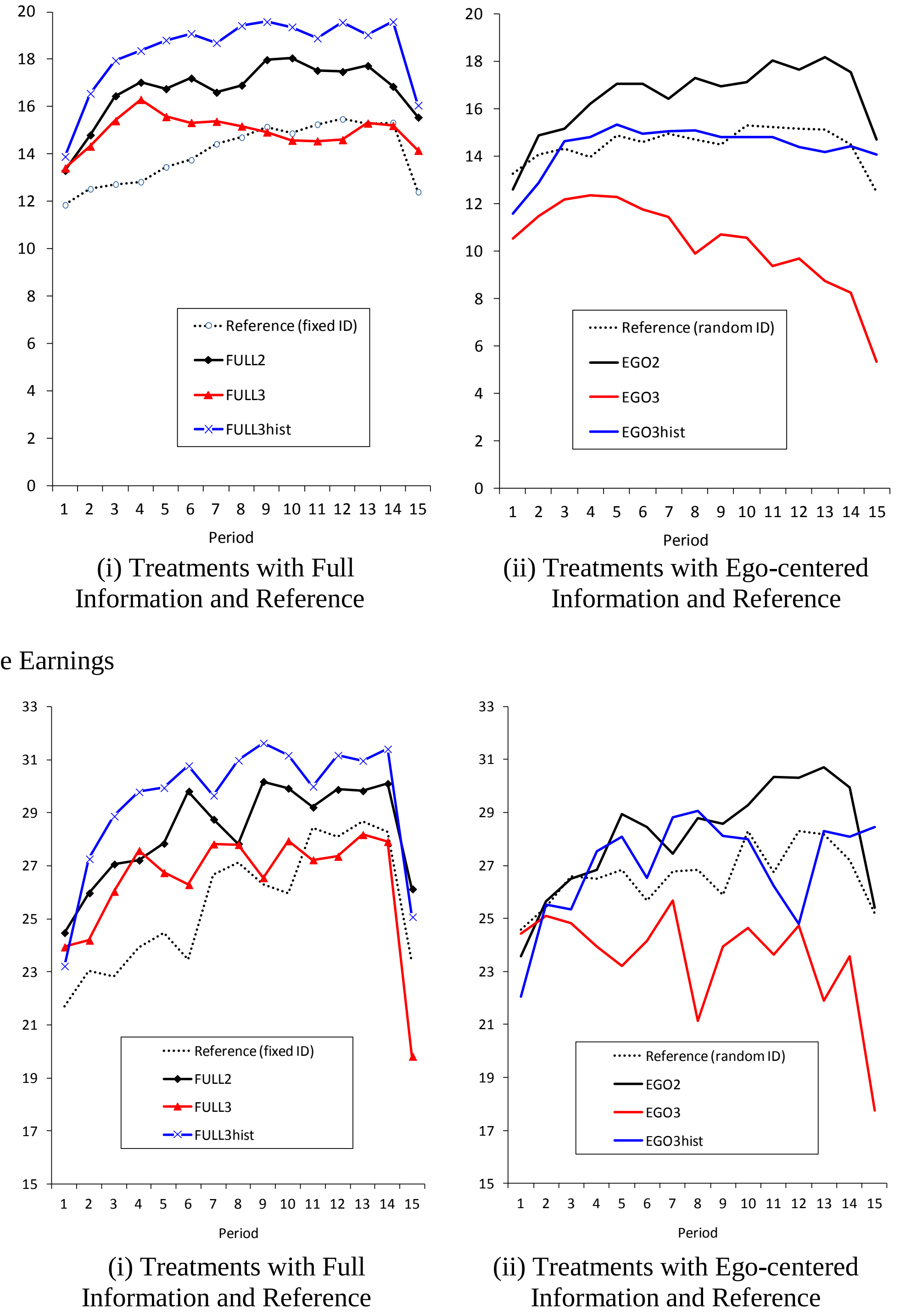
(a) Two Stage Periods (Reference (random ID), Reference (fixed ID), EGO2, FULL2)

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

* Note that information available to subjects at punishment stages varies by treatment. For an overview, see Table 2.

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

Note: All treatments in the left panels have fixed ID, whereas those in the right panels have randomly changing IDs.

Table 1. Summary of treatments, contributions and earnings

Treatment	Information Structure[1]	The number of stages in each period	History[2]	Informed higher order punishment opportunities	Total number of sessions	Total number of groups[3]	Average contributions	Average Earnings
(a) Treatments with Full Information and Reference (fixed ID)[4]								
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 Reference (random ID)[4]								
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. [4] All treatments under (a) have fixed ID while all treatments under (b) have randomly changing IDs.

Table 2. Information and Punishment Opportunities Available to Subjects in each Treatment

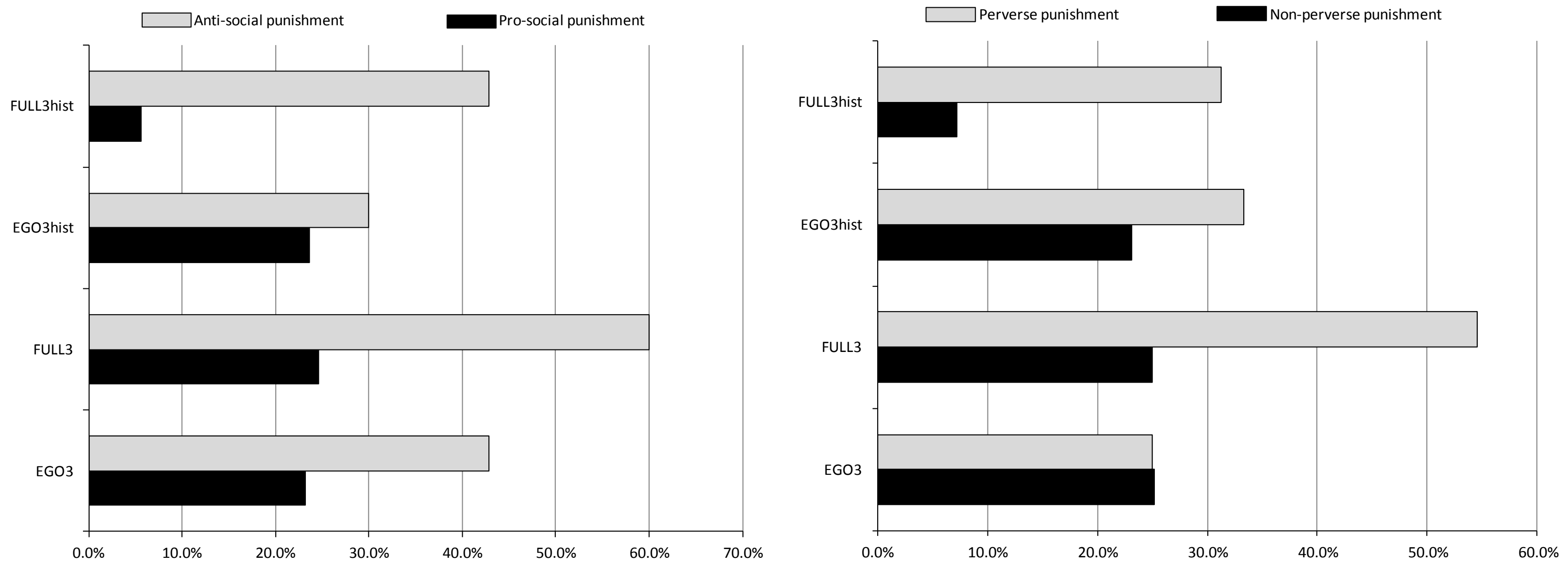
Treatment	Stage 1: Contribution Stage in Period *t*	Stage 2: First Punishment Stage in Period *t*	Stage 3: Second Punishment Stage in Period *t*	Who Subjects are Permitted to Punish in Stage 3
Reference (random ID)	No Information	Contribution decisions in period *t*	N.A.	N.A.
EGO2	No Information	(1) 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)).

Fig. 3. Percentage of 2nd stage punishment event that are counter-punished in 3rd stage, by category



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

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

Notes: [#1] We call a 2nd stage (i.e., 1st order) punishment event "pro-social" if it is directed at one who contributed less than the punisher. We call it "anti-social" if it is directed at one who contributed more than or equal to the punisher.

[#2] We call a 2nd stage punishment "non-perverse" (or normal) if it is directed at one who contributed less than the group's average contribution in the period. We call it "perverse" if it is directed at one who contributed more than or equal to the group's average contribution.

The length of each bar indicates the percentage of the relevant 2nd stage punishment events from *i* to *j* that are followed by 3rd stage punishment from *j* to *i*.

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 [1]	FULL2 [2]
(i) Average contribution in group in period *t*	-0.028*** (0.037)	-0.083 (0.068)
(ii) Max $\{(C_{it} – C_{jt}), 0\}$ [abs. neg. deviation in *t*]	0.27*** (0.029)	0.30*** (0.034)
(iii) Max $\{(C_{jt} – C_{it}), 0\}$ [pos. deviation in *t*]	0.072** (0.032)	0.084** (0.038)
(iv) Max $\{(C_{i(t-1)} – C_{j(t-1)}), 0\}$ [abs. neg. deviation in *t* - 1]	----	0.032 (0.030)
(v) Max $\{(C_{j(t-1)} – C_{i(t-1)}), 0\}$ [pos. deviation in *t* - 1]	----	-0.029 (0.048)
(vi) $p_{ji(t-1)}$ if $C_{j(t-1)} > C_{i(t-1)}$, else 0	0.90*** (0.20)	0.71*** (0.26)
(vii) $p_{ji(t-1)}$ if $C_{j(t-1)} \leq C_{i(t-1)}$, else 0	1.14*** (0.29)	1.14** (0.47)
(viii) $\sum p_{jk(t-1)}$, all $k \neq i$ such that $C_{k(t-1)} \leq C_{i(t-1)}$	----	-0.41* (0.25)
(ix) $\sum p_{jk(t-1)}$, all $k \neq i$ such that $C_{k(t-1)} > C_{i(t-1)}$	----	0.043 (0.33)
(x) *t* [period]	-0.14*** (0.029)	-0.064* (0.033)
Constant	-2.73*** (0.70)	-2.89** (1.26)
# of Observations	1680	1680
Log likelihood	-588.9	-511.2
Wald Chi-squared	134.69	125.73
Prob > Wald Chi-squared	.000	.000
Chi-squared Test on (vi) = (vii)		
Chi-squared	0.54	0.68
p-value	0.4643	0.4098

Notes: Random effects Tobit Regressions. 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. The number of left- (right-) censored observations is 1528(1) in column [1] and 1549(0) in column [2]. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

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*

	EGO3	EGO3hist	FULL3	FULL3hist
Independent variable	(1)	(2)	(3)	(4)
(i) Average contribution in group	0.15	-0.10*	-0.070	-0.45***
in period t	(0.096)	(0.057)	(0.064)	(0.13)
(ii) Max $\{(C_{it} – C_{jt}), 0\}$	-0.038	-0.073	0.32***	0.13**
[abs. neg. deviation in t]	(0.26)	(0.084)	(0.048)	(0.057)
(iii) Max $\{(C_{jt} – C_{it}), 0\}$	-0.092	-0.16***	0.040	-0.24**
[pos. deviation in t]	(0.080)	(0.060)	(0.058)	(0.12)
(iv) p_{jit} if $C_{jt} > C_{it}$, else 0	-0.026	0.24	1.92***	1.70***
	(0.20)	(0.17)	(0.45)	(0.60)
(v) p_{jit} if $C_{jt} \leq C_{it}$, else 0	0.078	0.056	0.82	3.61***
	(0.26)	(0.25)	(1.31)	(0.87)
(vi) $\sum p_{jkt}$, all $k \neq i$ such that $C_{kt} \leq C_{it}$	----	----	-0.61	-0.47
			(0.47)	(0.37)
(vii) $\sum p_{jkt}$, all $k \neq i$ such that $C_{kt} > C_{it}$	----	----	0.48	-0.91
			(0.78)	(1.09)
(viii) Period	-0.029	-0.0027	0.058	0.14***
	(0.094)	(0.066)	(0.044)	(0.051)
Constant	-2.74*	0.94	-6.95***	-0.0022
	(1.61)	(1.07)	(1.30)	(2.25)
# of Observations	143	120	1800	1620
Log likelihood	-131.99	-94.78	-623.07	-187.56
Wald Chi-squared	4.02	10.93	73.61	37.40
Prob > Chi-squared	.6746	.0906	.000	.000
Chi-squared Test on (vi) = (v)				
Chi-squared	0.04	0.39	0.70	4.07
p-value (2-sided)	.8421	0.5311	.4011	.0438**

Notes: Random effects Tobit Regressions. 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 3^{rd} 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 .05 level and at the .01 level, respectively.