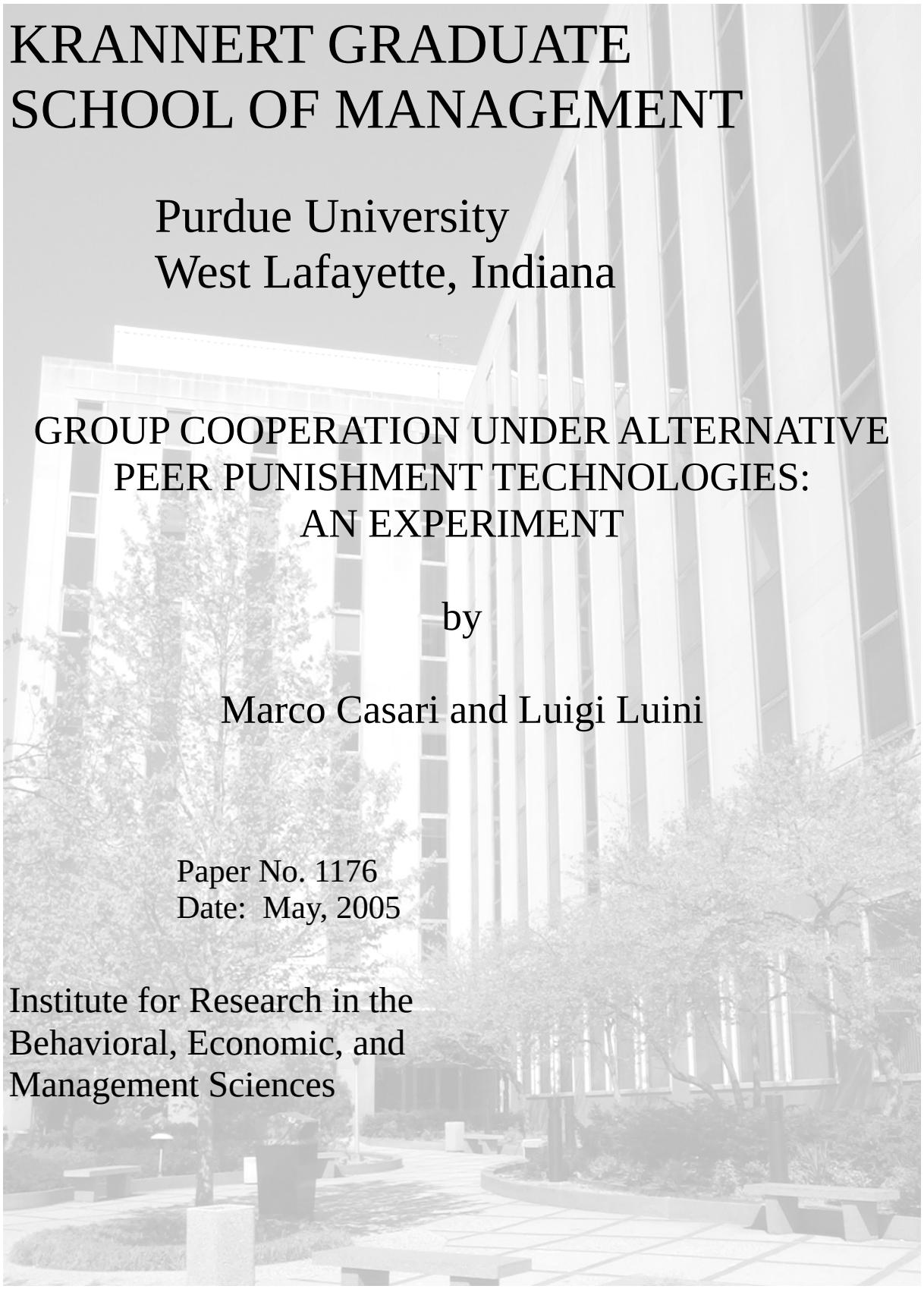


KRANNERT GRADUATE SCHOOL OF MANAGEMENT

Purdue University
West Lafayette, Indiana

GROUP COOPERATION UNDER ALTERNATIVE PEER PUNISHMENT TECHNOLOGIES: AN EXPERIMENT

by

Marco Casari and Luigi Luini

Paper No. 1176
Date: May, 2005

Institute for Research in the Behavioral, Economic, and Management Sciences

GROUP COOPERATION UNDER ALTERNATIVE PEER PUNISHMENT TECHNOLOGIES: AN EXPERIMENT[*]

Marco Casari

Purdue University

and

Luigi Luini

Università di Siena

Abstract

This paper experimentally studies peer punishment under three alternative technologies. We find that the choice of peer punishment technology has a substantial impact on group performance. First, under a technology where at least two subjects in the group must agree before another group member can be punished, group cooperation and group net earnings are the highest. Second, outcomes are similar regardless of whether punishment choices are simultaneously or sequential. These results suggest that punishment is not perceived as a second-order public good but is instead an emotional reaction unresponsive to changes in the strategic environment *(JEL* C91, C92, D23).

Keywords: decentralized punishment, public goods, other-regarding preferences, team production, experiments.

---

[*] Corresponding author: Marco Casari, Krannert School of Management, Purdue University, 403 West State Street, West Lafayette, IN 47907, Tel: 765 494 3598, Fax: 765 496 1567, casari@purdue.edu; Luigi Luini, Dipartimento di Economia Politica ,Università di Siena, Piazza San Francesco D´Assisi 5, 53100 Siena, Italy, Tel: +39 0577 232 608, Fax: +39 0577 232661 luini@unisi.it. We thank Aurora Gallego, Nikos Georgantis, Régis Renault, Francesco Ricci, participants to seminars at Universitat Jaume I of Castellon, THEMA at Univ. Cergy-Pontoise of Paris, and ESA meeting in Erfurt and Amsterdam for helpful comments. Thank you also to Francesco Lomagistro for the programming and for the help in running experiments. The usual disclaimer applies. This research was financially supported by a grant of Monte dei Paschi di Siena and by a Marie Curie Fellowship of the European Commission.

While peer punishment has been shown to increase group cooperation, there is still open debate on how group members endogenously enforce cooperative norms and on what motives other than self-interest drive individuals to punish. In this paper we address both topics by experimentally comparing alternative technologies of peer punishment.

Peer punishment is widespread in many field environments. The clothing company Diesel has been tremendously successful in building a strong work ethic among its employees by using merciless peer sanctioning (Stella, 1996).[1] To avoid excessive harvesting, fishing communities use several forms of peer punishment, including vandalism of boats and fishing nets or denying loans (Ostrom, 1996). Mining communities also use peer punishment such as social ostracism to discourage others from breaking strikes over working conditions (Francis, 1985).

While peer punishment in field contexts takes many forms, the experimental economic literature has focused on the very specific punishment technology where group members *simultaneously* choose if and how much to punish each member of the group without ever knowing the punishment choices of others and with unlimited discretionality on punishment choices (Ostrom et al., 1992, Fehr and Gaechter, 2000, 2002, Sefton et al., 2002, Bochet et al., 2002, Masclet et al., 2003). Important implications have been drawn from this punishment structure, but there are several reasons to examine alternative peer punishment technologies. First, doing so provides a robustness check for existing studies. So far the only systematic exploration of the above peer punishment structure concerns changes in the relative cost of

[1] In an interview the CEO of Diesel Renzo Rosso describes how he is welcoming the new employees: "Look, here all the doors are open to you. You can climb the ladder or change task as much as you wish. But watch out: if you do not show competence in what you do, they all will walk over you. It will not be me to throw you out, but your very own peers," Stella (1996), pp.24-26, (our translation) Civil servants in some public administrations have equally strict norms but of the opposite sign, where shopping during office hours is allowed and hard working habits are discouraged.

punishment (Carpenter, 2002, Andreoni et al., 2003, Putterman and Anderson, 2003). Our paper alters the structure of the technology along multiple dimensions. Second, variations in the technology of peer punishment may affect group performance. For any given social norm, changing the punishment technology may lead to more or less group cooperation. If a manager can shape the way team members interact, it may be possible for the manager to increase total production without, or in alternative to, manipulating the workers' social norms. Third, alternative peer punishment technologies may help to disentangle competing theoretical explanations regarding non-selfish motives to punish.

Our experiment includes three punishment technologies, "Baseline," "Sequential" and "Consensual." The Baseline technology implements a design similar to what is common in the literature. When agents are motivated purely by personal monetary earnings, there should be no punishment and complete free riding in all treatments. We know from previous studies that this is not an accurate description of experimental results but scholars have not yet settled on a single explanation. There can be different ways in which a norm of cooperation is enforced.

We consider two classes of motivations that generate distinct predictions about the pattern of punishment actions depending on whether the punishment choices take place simultaneously (Baseline) or sequentially (Sequential). On one hand, punishment could be an instrument that agents employ to achieve a desired distribution of earnings across group members. In that case, an "instrumental punisher" that is willing to punish a non-cooperator if no one else does, will happily free-ride *on punishment* if she knows that others will punish. On the other hand, punishment could be an expression of the emotional response to a norm violation that is independent of others group members punishment decisions. In this case, an "expressive

punisher" that is willing to punish a non-cooperator, will punish equally in both the Baseline and Sequential treatments. In this study we find a considerable amount of expressive punishment.

Each individual may have norms, i.e. standards of behavior about how individual group members ought to behave in a given situation. Some neurological evidence suggests that sanctioning a norm violator may be a source of utility for the punisher (de Quervain et al., 2004). One issue is that a group may face multiple norms, possibly conflicting ones.[2] In the Baseline technology all requests to punish are carried out and the group outcome results from the mixed impact of all individual norms. In the Consensual technology requests to punish are carried out only when there is a minimum coalition of two agents that shares a norm, i.e. want to punish the same agent. In either case, norms are not imposed by the experimenter but are endogenously determined by the group members. To some extent, the Consensual technology resembles to a legal system where laws are chosen or enforced only through social agreement.[3]

We measure performance of a punishment technology by its impact on group cooperation levels and group net earnings. The latter considers group earnings minus the costs of giving and receiving punishment. Group cooperation is measured using voluntary contributions to the production of a public good. Among the three treatments studied, the Consensual treatment provides the highest level of group cooperation and the highest group net earnings. Compared to the treatment when agents are not able to punish, the baseline and sequential treatments improved the level of group cooperation but yielded lower group net earnings. This is similar to the situation of a workplace that offers a high nominal wage, but employees do not stay because of the conflicts between workers.

[2] The two agents must agree on the target but can still disagree on the amount of the punishment. Consider forming a mixed team with civil servants and Diesel employees.

[3] A biblical rule: "One witness is not enough to convict a man accused of any crime or offense he may have committed. A matter must be established by the testimony of two or three witnesses." (Deuteronomy 19:15)

The rest of the paper is outlined as follow. In Section I we describe the experimental design. In Section II we present the theoretical predictions. The results are presented in Section III, and the conclusions follow in Section IV.

I. The Experimental Design

A. *Basic Design*

Our design consists of a public good experiment with three treatments of differing punishment technologies.[4] There are N=20 participants in each session. In every period the participants are randomly partitioned into four groups of n=5 individuals. In all treatments subjects participate for twenty periods in a finitely repeated public good game with and without punishment opportunities. In the first ten periods there is no punishment opportunity while in the last ten periods there is. Punishment opportunities are structured in three different ways: Baseline, Consensual, and Sequential (Table 1).

In the Baseline treatment, once group members are informed about each members' contribution to the public good, all punishment requests are simultaneously submitted. At a private cost of one token per punishment point, an agent can decrease the earnings of any other individual in her group by three tokens. In the case an agent receives punishment points from two or more agents, her earnings reduction is the cumulative effect of all requests. Punishment on a targeted agent is carried out irrespectively of the number of requests. This is a common protocol in the experimental literature, adopted for instance by Fehr and Gaechter (2000).

In the Consensual treatment, participants simultaneously place their punishment requests. If a subject is the target of the punishment request of just one other subject, that punishment request

[4] The Instructions for the Consensual treatment can be found in Appendix B. The whole experiment was framed in neutral terms.

is ignored. When at least two group members request to punish a subject, their punishment requests are carried out. Subjects are informed of the outcome of their requests.

Finally, requests to punish in the Sequential treatment are not placed simultaneously but in (n – 1) steps. All requests are carried out, but every participant considers punishing each one of the other (n – 1) group members in separate steps. In step one, a subject knows that the other (n – 2) group members will have the opportunity to punish this agent after her. In step two, a subject knows that one other group member has already had the opportunity to punish the same individual and also knows the cumulative amount of punishment already inflicted. And so on for the remaining steps.

The treatment without punishment opportunity serves as a control for the treatment with punishment opportunity. The same n subjects interact ten periods without punishment opportunities and then ten periods with the opportunity to punish.[5]

B. Payoffs

In the treatments without punishment, in each period each of the n subjects in a group receives an endowment of y tokens. A subject can either keep these tokens for herself or invest g_i tokens ($0 \leq g_i \leq y$) into a project. The decisions about g_i are made simultaneously. The period monetary payoff for each subject i in the group is given by

$$\pi_i^1 = y - g_i + a \sum_{j=1}^{n} g_j \qquad (1)$$

where a is the marginal per capita return from a contribution to the public good, $1/n < a < 1$. The total payoff from the no–punishment condition is the sum of the period–payoffs, as given in (1), over all ten periods. Note that (1) implies that full free–riding ($g_i = 0$) is a dominant strategy in

[5] We run three additional sessions where the ten periods with punishment opportunities were placed before the ten periods without the punishment opportunity. These results are not reported here.

the stage game. This follows from $\partial \pi_i^1 / \partial g_i = -1 + a < 0$. However, the group payoff $\sum_{i=1}^{n} \pi_i^1$ is maximized if each group member fully cooperates ($g_i = y$) because $\partial \sum_{i=1}^{n} \pi_i^1 / \partial g_i = -1 + na > 0$.

The major difference between the no–punishment and the punishment conditions is the addition of a second decision stage after the Simultaneous contribution decision in each period. At the second stage, subjects are given the opportunity to punish each other after they are informed about the individual contribution of the other group members. Group member *j* can punish group member *i* by assigning so–called punishment points p^i_j to *i*. There are three different treatments for the part with punishment opportunities.

In the Baseline (Simultaneous) treatment for each punishment point assigned to *i* the first–stage payoff of *i*, π_i^1, is always reduced by three tokens. Agent *i* takes all punishment decisions at once $\{ p^1{}_i, \ldots, p^{i-1}{}_i, p^{i+1}{}_i, \ldots, p^n{}_i \}$, where $p^k{}_i \in \{0,1,\ldots,7\}$, and simultaneously with the other agents. For received punishment points, agent *i*'s payoff is reduced by $\sum_{k \neq i}^{n} e(p_k^i)$, where $e(p^i_j) = 3\, p^i_j$ is the effectiveness of punishment function. For punishment points given to others, agent *i*'s payoff is reduced by $\sum_{k \neq i}^{n} c(p_i^k)$, where $c(p^j{}_i) = p^j{}_i$ is the cost of punishment. This design has the important feature of holding the fine-to-fee ratio $e(p^i{}_k)/ c(p^i{}_k)$ constant – and equal to 3 – in order not to alter the "price" of punishment.[6] The pecuniary payoff of subject *i* from both stages, π_i, can therefore be written as:

$$\pi_i = \pi_i^1 - \sum_{k \neq i} e\left(p_k^i\right) - \sum_{k \neq i} c\left(p_i^k\right) \qquad (2)$$

[6] The reason is to avoid confounding effects in the interpretation of results due to differential "pricing" of punishment. For the same reason we allowed period earnings of a subject to be negative. Not doing so would have increased the fine-to-fee ratio of the marginal punisher. In the experiment, negative period earnings where infrequent. When ignoring the punishment given to others, it amounts to 3.3% of the observations with punishment opportunities. Cumulative earnings were always positive.

The total payoff from the punishment condition is the sum of the period–payoffs, as given in (2), over all ten periods.

In the Consensual treatment, punishment is Simultaneous and employs the same cost and effectiveness functions as before. However, an agent is punished only if at least two agents requested it. Hence, the payoff function for both stages is:

$$\pi_i = \pi_i^1 - K(i)\sum_{k\neq i} e\left(p_k^i\right) - \sum_{k\neq i} K(k)c\left(p_i^k\right) \qquad (3)$$

where $K(i) = \left(\sum_k I_{\{i,k\}}\right) \geq 2$. The function $I(i,k)$ equals one when agent k requests to punish agent i, $p^i_k > 0$, and equals zero otherwise. In practice, only a coalition of 40% of group members or larger is allowed to punish a member. Isolated requests to punish agent *i* have no effect and no cost is charged for that request. If the punishment request is not carried out then the requesting subject is informed about it and the targeted subject will not know of such request.

In the Sequential treatment the payoffs are given by (2) like in the Baseline treatment but the timing of decisions is different. Instead of being taken all at once, punishment decisions $\{ p^1_i,\ldots, p^{i-1}_i, p^{i+1}_i,\ldots, p^n_i\}$ are broken down into (n–1) distinct steps where at step k agent i makes a single decision $p^{j(k)}_i$. The order of punishment decisions $j(k)$ is random and such that within the period agent i has an opportunity to target all other agents in the group, $\{j(1),\ldots, j(n-1)\}=\{1,\ldots, i-1, i+1,\ldots,n\}$. After each step, there is an update on the cumulative punishment received by each agent in the group.

C. Parameters and Information Conditions

The experiment is conducted in a computerized laboratory where subjects anonymously interact with each other.[7] No subject is ever informed about the identity of the other group

[7] For conducting the experiments we used the experimental software "z-Tree" developed by Urs Fischbacher (1998).

members. No communication among subjects was allowed. In all treatment conditions the endowment is given by $y = 20$, groups are of size $n = 5$, the marginal payoff of the public good is fixed at $a = 0.4$, and the number of participants in a session is $N = 20$. In each period subject i can assign up to seven punishment points p_i^j to each group member j, with $j \neq i$ irrespective of their first stage earnings. In all treatment conditions subjects are publicly informed that the condition lasts *exactly* for ten periods. When subjects play the no–punishment opportunity condition they know that a session consists of two conditions but do not know the rules for the second condition. After period ten of the first condition in a session, they are informed that there will be a "new experiment" and that this experiment will again last exactly for ten periods. They are also informed that the experiment will then be definitely finished.[8]

In the no–punishment conditions the payoff function (1) and the parameter values of y, n, N, and a are common knowledge. At the end of each period subjects in each group are informed about the total contribution Σg_j to the project in their group.

In the punishment conditions the payoff function (2) or (3), in addition to y, n, N, a, and the protocol of the punishment requests are common knowledge. Furthermore, after the contribution stage subjects are also informed about the whole vector of individual contributions in their group. To prevent the possibility of individual reputation formation across periods each subject's own contribution is always listed in the first column of his or her computer screen and the remaining four subjects' contributions are *randomly* listed in the second, third, fourth, or fifth column respectively. Thus, subject i does not have the information to construct a link between individual contributions of subject j across periods. Therefore, subject j cannot develop a reputation for a particular individual contribution behavior. This design feature also rules out that

[8] Each condition was preceded by a trial period to familiarize the subjects with the software.

i punishes *j* in period *t* for contribution decisions taken in period *t'* < *t*. Subjects know their own punishment activities, the aggregate punishments imposed on them by the other group members, and the *aggregate* punishment imposed on *other* group members.[9]

In the Sequential punishment treatment subjects know the step where they are at and the cumulative *aggregate* punishment imposed on *other* group members up to the previous step. Hence, they receive more detailed information about punishment than in the Baseline treatment because they can see both the end-of-period sum and some disaggregated statistics about the individual components of this sum. However, they are not informed however about the amount of punishment they have personally received until the end of the period. In all treatments, subjects are also not informed about the *individual* punishment requests of the other group members. Both provisions are meant to prevent, as much as possible, a subject from using punishment to pay-back others for their requested punishments.

II. Predictions

We outline the predictions for the experimental condition with punishment opportunities. In particular, we provide the intuition for the punishment stage predictions of three alternative models, canonical, expressive and instrumental. Appendix A provides more details and the proofs. All predictions are made under the following assumptions: (a) one-shot interaction; (b) common knowledge about agents' preferences; and (c) risk-neutrality. In the experiment the probability that an agent was re-matched with the same four people was less than 2 percent. Duffy and Ochs (2004) have shown that using a similar random matching protocol in an

[9] This provision can make a difference when subjects do not know the preferences of others. When a subject can only observe the punishment points she gave or received (Fehr and Gaechter, 2000), learning about these preferences may be slower than here. In our setting, a subject can see if a social norm was enforced with respect to any other subject in her group.

experiment of *indefinite* length does not induce cooperation levels higher than the one-shot prediction.

Consider the decision of agent *i* to punish agent *j* at the end of the contribution stage. Agent *i* knows the contribution levels, g_k, of all agents, including agent *j* and must choose a number of punishment points $p^j_i \in \{0, 1, 2, 3, \ldots\}$ where zero means no punishment. We assume that the utility function of agent *i* is the following:

$$u_i (g_1, g_2,..., g_n; p^j_1, p^j_2,..., p^j_n) = \pi^1_i (g_1, g_2,..., g_n) - c(p^j_i) + v_i (p^j_1, p^j_2,..., p^j_n) \qquad (4)$$

There are three components: first stage earnings, π^1_i, punishment cost, *c*, and utility from punishment, v_i. While first stage earnings depend on everybody's contribution levels, g_k, punishment cost depends on the points of punishment agent *i* has given to agent *j*, p^j_i. The punishment of agent *j* may increase the utility of agent *i* by v_i.[10]

The three models considered in this paper differ only in the utility from punishment, v_i:

- Canonical, $v_i = 0$ (5)
- Expressive, $v_i = v_i (p^j_i)$ (6)
- Instrumental, $v_i = v_i (p^j_i, p^j_{-i})$ (7)

We denote with $p^j_{-i} = \sum^n_{k \neq i} p^j_k$ the sum of the punishment points that others have given to agent *j*, and assume that $v_i'(p^j_i) = v_i'(p^j_{-i}) > 0$ and $v_i''(p^j_i) = v_i''(p^j_{-i}) < 0$.

If subjects apply the backward induction logic, the canonical equilibrium prediction in all three treatments is that all subjects will contribute nothing to the public good and will punish nothing. In fact, choosing $p^j_i > 0$ is a monetary cost that does not generate any monetary benefit in a one-shot interaction.

[10] Note that the function v_i can vary for a different target agent *j*, a different vector of first-stage contributions (g_1, ..., g_n), and a different effectiveness function $e(p^j_i)$, which sets the fine-to-fee ratio.

To understand expressive and instrumental motives for punishment consider the example illustrated in Figure 1. Agent *i*'s utility increases in the level of punishment given to agent *j*, $v_i' > 0$, but the utility gain is smaller for higher punishment levels, $v_i'' < 0$. Punishment may well be positive but the actual number of punishment points chosen depends on the private cost of punishment:

$$\bar{p}_i^j = \underset{p_i^j \in \{0,1,2,\ldots\}}{\arg\max} \left\{ u_i\left(\pi_i^1,\ p_i^j,\ 0\right) \right\} \qquad (8)$$

As the following discussion focuses on a given target agent *j*, for simplicity we use $\bar{p}_i$ instead of $\bar{p}_i^j$. Both models predict that increasing the cost of punishment *c*, lowers $\bar{p}_i$. This "price effect" of punishment that has been found by several experimental studies (Carpenter, 2002, Andreoni et al., 2003, Putterman and Anderson, 2003). The basic difference between the two models is that what matters for an expressive agent *i* is only the punishment that she *personally* carries out, p_i while an instrumental agent *i* equally values the punishment she gives and the punishment that others give. As a consequence, for an instrumental punisher, $\bar{p}_i$ is the "standalone punishment level" (when nobody else punishes, $p_{-i}=0$) and is the upper bound to what she will do. The actual number of punishment points requested by an instrumental punisher rely upon strategic considerations based on what other agents choose in reference to agent *j*. On the contrary, for an expressive punisher $\bar{p}_i$ is always the optimal choice.

An expressive punisher's punishment choice is independent of any strategic considerations. In equilibrium all expressive punishers with $\bar{p}_i > 0$ will request to punish agent *j*. Hence there may be multiple requests to punish agent *j* and the total punishment agent *j* is $\sum_{k \neq j}^{n} e\left(\bar{p}_k^j\right)$. The

expressive model predicts an equal total punishment for agent *j* under the Baseline or Sequential technologies, and less than or equal total punishment under the Consensual technology.

For an instrumental punisher the essential issue is the total impact on agent *j*, and she has no objections to others doing the "dirty job" of punishing. She actually prefers it because it saves her the punishment cost. This framework was adapted from the model that Varian (1994) developed for voluntary public good contributions. The equilibrium strategy in the Baseline technology is *for the agent with the maximum standalone punishment level to request the punishment,* $\bar{p}_i = \max_{k \neq j} \{ \bar{p}_k \}$*, and for all the other agents to free ride on the punishment*. In the Consensual technology, two agents, one of these being the agent with the maximum standalone punishment level, will punish agent *j* in equilibrium. The total punishment of agent *j* in the Consensual treatment will be less than or equal to the Baseline treatment. The equilibrium strategy in the Sequential technology depends on the order of move in the punishment phase. The intuition is as follow. Suppose everybody wants to punish agent *j*, $\bar{p}_k > 0$ for all k≠j. If agent *i* is the last mover and agent *j* has yet to be punished by other group members, than agent *i* will choose to punish with level $\bar{p}_i$. If another agent *h* is not the last mover, she will choose to punish agent *j* only if her standalone punishment level is *much* higher than agent *i*'s, $\bar{p}_h >> \bar{p}_i$. If agent *h* chooses to punish, agent *i*, the last mover, will not punish agent *j*. In equilibrium, when agents know the order of moves, punishment in the Sequential technology is carried out by only one agent and total punishment is less than or equal to that in the Baseline technology. While the expressive model predicts no punishment differences across steps with the Sequential technology, the instrumental model predicts that the burden of punishment falls disproportionally on the agent that moves last.

III. Results

A total of 240 subjects were recruited among the general undergraduate student population of the University of Siena via ads posted around campus asking to email or call. No subject had participated in this type of experiment before, and each subject participated in only one of the experimental sessions. Twelve sessions were conducted between March and October 2003. Each session lasted between 1 hour and 50 minutes and 2 hours and 30 minutes. Payment was done privately in cash at the end of each session and totaled 12.40 euros per subject on average.[11]

The results are grouped into two sub-sections, one referring to aggregate cooperation and net payoff (Results 1-3) and one concerning individual decisions to punish (Results 4-7).

A. Aggregate cooperation and surplus

When subjects have the opportunity to punish, contributions to the public good increase (Result 1) and stay high over time (Result 2). In addition to replicating these well known results, we also find sharp differences regarding group cooperation levels and group net earnings according to the punishment technology used. In particular, group cooperation (contribution) and group net earnings are highest in the Consensual treatment (Result 3).

RESULT 1: *The existence of punishment opportunities causes a rise in the average contribution level from 17% to 29% of the endowment. In particular, while the average contribution raises in all treatments, the rise is largest in the Consensual treatment.*

RESULT 2: *In the no-punishment condition average contributions converge over time close to full free riding. In contrast, in the punishment condition average contributions are stable or*

[11] At the October 2003 rate of $1.17 per euro, it is equivalent to $14.50. This amount includes the show up fee that was 3 euros for the four sessions conducted before October and 5 euros afterwards. The amounts in the instruction were quoted in "Tokens". A token was converted into 0.02 euros.

increasing over time. In particular there is a steady growth in contribution levels in the Consensual treatment.

Support for Results 1 and 2 comes from Table 2 and Figure 2. Without a punishment opportunity the average individual contribution across all treatments is 3.31 tokens. This average value hides a declining trend from 5.92 tokens in period one to 1.82 in period ten, which is similar in the three treatments. When the opportunity to punish is introduced, the average individual contribution across periods and treatments with punishment is 5.77. A nonparametric Wilcoxon signed ranks test shows that this difference in contributions is significant at the one percent significance level (p=0.0061). In the first period, with punishment opportunities (period eleven), there is a "jump" in the average contribution to 5.21 tokens that grows over time to 6.50 in period twenty. This jump in contribution between the last period without punishment and the first period with punishment is significant at a one percent level according to a Wilcoxon signed ranks test (p=0.0002).

Besides these common patterns, each punishment technology shows remarkable peculiarities. Overall contributions under a Consensual technology are substantially higher than in the other two (8.46 vs. 4.46 Baseline and 4.38 Sequential).[12] Moreover, while the time trend is increasing for the Consensual technology (period one-ten, 6.94-9.76), it is roughly stationary for the other two (4.01-5.65 Baseline, 4.62-4.10 Sequential). Such differences are summarized by the analysis of relative payoff gains with and without punishment in Result 3.

RESULT 3: *In all treatments punishment opportunities initially cause a relative payoff loss. The Consensual treatment is the only treatment in which relative payoff gains are found and this is only observed in the final four periods. In the final period of the Consensual treatment the*

[12] A non parametric Wilcoxon signed ranks test shows that the difference in contributions with and without sanction opportunities between the Consensual treatment on one side and the other two treatments on the other side is significant at a ten percent level (p=0.0768).

relative payoff gain is 13 percent. In the Baseline and Sequential treatments the relative payoff losses remain throughout all periods, although they become smaller over time. In the final period of the Baseline and Sequential treatments the relative payoff loss is roughly 20 percent.

Support for Result 3 comes from Table 2 and Figure 3. Normalizing the earnings in the final period of the no punishment condition to 100, then earnings in the first period with punishment are equal to 57 in the Baseline treatment, 53 in the Sequential, and 85 in the Consensual. By the end of the session, all of these values have increased. While the Baseline is at 80 and the Sequential is at 78, which are still below the reference value without punishment, the Consensual treatment is above, at 113. A Wilcoxon signed ranks test shows that the differences in group net earnings between the last period with and without sanction opportunities are significantly different in the Consensual treatment compared to the other two treatments at a five percent level (p=0.0364).

Hence, the Consensual punishment technology, by its ability to endogenously minimizing conflict while still maintaining incentives to cooperate, clearly dominates the Baseline and Sequential technologies in terms of group contribution levels and group net earnings. Let us define a punishment rate as the average number of punishment points assigned to a particular contribution action, $\left(\sum_j^n \sum_{k\neq j}^n p_k^j\right)/n$ for Baseline and Sequential and $\left(\sum_j^n K(j) \sum_{k\neq j}^n p_k^j\right)/n$ for Consensual. The punishment rate was 1.70 in the Consensual compared to 2.47 in the other two treatments (Figure 4). This difference persists also after adjusting for the variations in group contribution across treatments (Table 3). For any given contribution level, lower punishment rates translate into a smaller deadweight loss. One reason for the lower punishment rate is that all punishment requests made by just one agent were

ignored. Had those requests not been ignored, the punishment rate in the Consensual treatment would have been 29.5 percent higher (full sample, Figure 4). Interestingly, while less than one out of every ten requests to target full free-riders was censored, about three out of four attempts to punish strong cooperators with contributions (15,20] were blocked. The Consensual technology endogenously filtered out the anti-social norm of a minority that was targeting cooperators, thus enhancing the incentives to cooperate. The Baseline and Sequential technology instead allowed a minority to freely harm strong cooperators and hence group incentives for cooperation.

To provide additional statistical evidence for this explanation and to facilitate the comparison with Fehr and Gaechter (2000) we also present a regression analysis of punishment behavior. As a complement to the use of *absolute* contribution levels employed in Table 3, this analysis also captures the effect of punishment on subjects' *relative* contributions in respect to the group average. Table 6 contains a model and an ordinary least-squared (OLS) regression where the dependent variable is "received punishment points" of a subject and the independent variables comprise "strong cooperator", "others' average contribution", "positive deviation" and "absolute negative deviation", respectively. The latter variable is the absolute value of the actual deviation of a subject's contribution from the others' average in case that his or her own contribution is below the average. This variable is zero if the subject's own contribution is equal or above the others' average. The variable "positive deviation" is constructed analogously. The variable "strong cooperator" is one if the subject's contribution is above fifteen tokens and zero otherwise. This variable retains the absolute scale of the contribution level and may capture the tendency, mentioned above, to target highly cooperative subjects. The model also includes period and session dummies. In all treatments, the coefficient of the "absolute negative

deviation" is positive and significant at the one percent level. This result reinforces the conclusion that free riders can reduce the received punishment by increasing their contributions. Although the positive coefficient is not significant, in the Baseline and Sequential treatment, strong cooperators were targeted for punishment. However, in the Consensual treatment, strong cooperators were less likely to receive punishment (significant only at the 10 percent level). This result holds when controlling for the relative contribution with respect to the group.

What stands out in the analysis of group cooperation levels across treatments is the superiority of the Consensual technology. This technology realized a relative payoff gain through a contribution level 90 percent higher than the Baseline treatment and 10 percent lower punishment costs (Result 3).

B. Motivations to punish

We now turn to a comparison of the patterns of individual punishment decisions with the predictions of the models of instrumental and expressive punishment (Results 4-7).

RESULT 4: *In the Baseline treatment, approximately half of the times that a subject is punished, two or more subjects have requested the punishment.*

Support for Result 4 can be found in Table 4. Such a high frequency of multiple requests to punish the same actions can be better explained by a model of expressive punishment than by a model of instrumental punishment. We now discuss, within the model of instrumental punishment, possible reasons to expect multiple requests to punish the same agent. As none of them is explaining the magnitude of Result 4 while the model of expressive punishment does, we conclude that the latter model is more accurate.

Trembling hand. As a preliminary exercise, we consider the instance that the actual number of punishers is simply the outcome of money maximizing agents, as in the canonical model, with a tendency to make random mistakes in the decisions to punish or not any one of the other four agents. By chance there can be none, one, two, three, or four requests to punish an action, *k*. However, the empirical distribution over the frequency of punishment (Table 4) is statistically different from the distribution of draws from a binomial distribution. Assuming that punishment decisions are drawn, a binomial distribution, $\binom{n}{k}p^k(1\text{-}p)^{n-k}$, where n=4 and p is calibrated to fit the share of actions not punished (p=0.25 yields Pr{k=0}=0.316), predicts that many actions should be punished by just one agent, Pr{k=1}=0.42, substantially more than what one actually observes.[13] Moreover, if punishment choices were random one would not expect to find that free riders are a more frequent target than cooperators (Table 3). Hence, a trembling hand cannot explain the high frequency of punishments carried out by two or more agents. On the contrary, it reinforces the need for an alternative explanation.

Preferences for heavy punishment. There is an upper bound of seven punishment points that a single agent can request. A maximum request has a considerable impact on the earnings of an agent, namely a reduction between 40% and 105%.[14] Yet, if a subject wants to punish more, $\bar{p}_i^j$>7, and there is another similar punisher in her group, in equilibrium there are multiple requests to punish. Such an event can be fully ruled out only by removing the upper bound to

[13] A nonparametric Chi-squared test shows that the predicted punishment events done by a binomial distribution with p=0.25 for the cases of three and four (5.08%), two (21.09%), and one (31.64%) requests are different from the observed ones at all conventional significance levels (p<0.0001) for each one of the Baseline sessions. To carry out this test one must assume that each period yields an independent observation.

[14] Seven points of punishment reduces earnings by 21 tokens. If everybody free rides, g_i=0 for all i=1,2,3,4 then π_i^1=20. If one agent free rides, g_1=0, and all others fully cooperate, g_i=20 for i=2,3,4, then π_1^1=52 and π_i^1=32 for i=2,3,4.

punishment in the experimental design. Given the design, one measure of the extent of the censoring occurred is the proportion of seven-point requests, which is a modest 5.5 percent (Figure 5). Moreover, the proportion of subjects whose cumulative punishment is above seven points is 8.2 percent (8.0 percent in the Sequential treatment). We conclude that the presence on an upper bound to individual requests to punish may explain only a small fraction of the multiplicity of punishment requests.

Ignorance of others' willingness to punish. Consider a situation where agents differ in their standalone punishment level, $\overline{p}_i^j$. A subject knows her type and has a belief about the type distribution in the general population but does not exactly know who is currently in her group. This bayesian version of the game has some appeal, especially given the anonymity of the experimental setting.[15] A subject who believes that she is in the lower tail of the distribution will not request to punish while a subject who believes to be in the upper tail of the distribution may. This conjecture implies that multiple punishment requests toward the same subject would be relatively close in amount in comparison to the average punishment request across different subjects.[16] A possible test of this conjecture comes from decomposing the variance of individual request amounts p^j_i as $Var[p^j_i] = Var_x[E[p^j_i|j]] + E_x[Var[p^j_i|j]]$, where j is a targeted agent. Variance decomposition in Table 5 shows $E_x[Var[p^j_i|j]] > Var_x[E[p^j_i|j]]$, which suggests – contrary to this conjecture – that there is more diversity in the level of punishment among

[15] Yet, one could argue that group members' attitude toward punishment could reliably be inferred from the profile of contribution levels if cooperators tend to sanction more and tend to target free-riders. That could substantially reduce the uncertainty about others' punishment preferences and hence the impact of this type of explanation for multiple requests to punish.

[16] As a side note, if a subject updates over time her distribution over population types as she observes other subjects' punishment choices that might reduce the multiplicity of punishment requests over time. Any empirical statistics suffers from not having contribution conditions constant over time. Of all punishment decisions in the first half of Baseline sessions, about 42.3% were carried out by one request and this figure raises to 54.4% in the second half.

multiple requests toward the same subject than among average punishment requests toward different subjects.

Multiple equilibria. The model of instrumental punishment has multiple Nash equilibria when the highest punishers have very similar preferences for punishment.[17] In some of these equilibria there are multiple punishment requests. In order to account for the high empirical frequency of this event, one has to assume that this rather peculiar situation is very common in the experiment. Namely, that the set of top punishers share not only a generic social norm that classifies behavior into punishable or not punishable but also agree on the exact amount of punishment that each action deserves and have a similar preference for enforcing that norm.[18] A possible way to test this conjecture is to look at the Sequential treatment, where the multiplicity of equilibria disappears and so a reduction in the multiplicity of punishment requests should be observed. This reduction should become more pronounced as the proportion of top punishers increases. The similarity between the empirical distributions of the number of punishment requests in the Sequential and Baseline treatments (Result 6) suggests that the multiplicity of equilibria explanation can be rejected.

RESULT 5: *In the Consensual treatment, 44 percent of the times that a subject is targeted by punishment requests, the punishment is not carried out because only one subject requested the*

[17] See the Appendix with the case with |M|>1, BAS 3I and SEQ 2E.

[18] Suppose that these subjects have not yet developed a norm on how to coordinate while punishing and they randomize among possible equilibrium. A rough estimate suggests that around three quarters of the decisions to punish should fall in the situation above if the multiple equilibria explanation is calibrated using the data in Table 4. Given |M|=2 and a maximum $\bar{p}_i^j$ =3 for all i, if subjects randomize uniformly across {0,1,2,3} punishment points, then there are 16 possible outcome (1 of no punishment, 6 of one request, 9 of two requests). We set to 100 the total of one or two requests to punish. Empirically 47.4% of actions have two requests. That requires that of all decisional situations where there is punishment, 75.8% are with |M|=2 as above and the other 24.2% are with |M|=1. Computations done for $\bar{p}_i^j$ >3 yields a frequency higher than 75.8%. Correcting computations by including situations with |M|=3,4 will lower the 75.8% frequency but not substantially.

punishment. A relatively high portion of the requests that are filtered out are directed at punishing cooperators.

RESULT 6: *In the Sequential treatment, about half of the times a subject is punished, two or more subjects have requested the punishment. The empirical distribution of requests for punishment is rather similar to the one in the Baseline treatment.*

Support for Result 6 comes from Table 4. Any comparison between Sequential and Baseline distributions of request of punishment is only suggestive because it relies on the assumption of an identical underlying contribution pattern. There are differences between the two but the aggregate level of contributions with and without punishment have some similarities (Figure 2). If the multiple requests to punish originate on a failure to coordinate among subjects, and based on a conjecture that they may be somewhat comparable, one would have expected a substantial increase in punishment carried out by only one subject. Instead, there is only a very small increase. The persistence of such a high frequency of multiple requests to punish the same actions lends more support to a model of expressive punishment than to a model of instrumental punishment.

RESULT 7: *In the Sequential treatment, more punishment is requested in earlier steps than in later steps. More precisely, while the amounts requested by punishers are similar for all the steps, the frequency with which contribution choices are punished is higher in steps one and two than in steps three and four.*

As shown in Figure 6, there is almost 40% more punishment in step one than in step four, which is at odds with both the instrumental and expressive model of punishment.[19] The pattern of Figure 6 can be more clearly interpreted when seen as the product of two components, a time profile of frequencies and a time profile of amounts. The latter component records the step-by-

[19] See predictions SEQ 5I and SEQ 3E in Appendix.

step average punishment points among *punishers only*. If a subject requests zero punishment that step, she is excluded from the computation of the time profile of amounts. The actual time profile of amounts is basically flat, i.e. when a subject decides to punish she chooses on average the same amount irrespective of the step. The driving force for the pattern in Figure 6 is the declining time profile of frequencies, i.e. the step-by-step average *fraction* of punishers on all subjects in the group. In each one of the Sequential sessions, more subjects actually request positive punishment in steps one and two than in steps three and four.

The declining time profile of frequencies contradicts both models, as the expressive punishment model predicts a flat profile while the instrumental punishment model predicts an increasing profile.

On the other hand, the flat time profile of amounts is exactly predicted by the expressive punishment model while under some assumptions it is compatible with the instrumental punishment model as well. While neither model fares perfectly, the distance between predictions and evidence is larger for the instrumental than for the expressive punishment model.

The predictions of the instrumental model rely on the assumption that agents take strategic considerations into account. If so, they should be able to backward induct and force the last agent to pay for punishing in the Sequential treatment. Let us consider a variation of the instrumental punishment model where agents are unable to backward induct. Such a model can accommodate the multiplicity of punishment requests in the Sequential treatment (Result 6) that was troublesome to explain when backward induction is assumed in that model. Regarding Result 7, this adjusted model correctly predicts a declining time profile of frequencies.[20] The flat time profile of punishment amounts, though, is a roadblock, as there is a prediction of a strictly declining pattern.

[20] It actually predicts a steeper decline in punishment frequencies than the one observed.

IV. Conclusions

Our experimental study of peer punishment technologies leads to three major conclusions.

First, the specific rules that govern punishment interaction do influence group performance in the voluntary provision of a public good (Result 1). In all treatments each agent had a costly opportunity to decrease the earnings of others in the absence of any personal material benefit. While this study replicates and confirms the robustness of the qualitative results in the literature (Ostrom et al., 1992, Fehr and Gaechter, 2000, 2002, Bochet et al., 2002), it also points to the significant impact of the punishment technology employed. An important result is that when punishment can be carried out only with the agreement of a coalition of agents, the group performs better than when each individual has full discretionality on imposing punishment on other group members. We measure performance both in terms of group contribution and group net earnings.

Second, peer punishment is costly and has the potential to destroy more resources than it generates, hence it does not automatically benefit the group (Result 3). Given the increases in cooperation levels when the opportunity to punish is offered, one may conclude that peer punishment is an appealing solution to the free riding problem. However, unless we specify the institution that governs peer punishment, this solution can have severe drawbacks. We find that in two out of the three treatments, the ability to punish others actually lowers group net earnings. This loss is especially pronounced during the periods of transition directly after the opportunity to punish is provided.

In line with most of the experimental literature on punishment, this study did not reveal the identity of the punisher nor permit counter-punishments, hence preventing revenge. Allowing for it would have further exacerbated group conflict and brought up aspects of peer punishment that are detrimental for group performance (Nikiforakis, 2004; Masclet et al., 2004). Anthropological studies of societies without a judicial system have pointed to the danger of the spontaneous human tendency to engage in peer punishment (Lowie, 1970, p.400, Girard, 1977, p.16-22). Our findings provide indirect support for the role of a legal system in the administration of punishment. Legal systems restrict sanctioning to the violation of shared rules and censor individual attempts to punish socially virtuous actions, hence channeling agents' punishment attitudes toward beneficial ends for society (Kosfeld and Riedl, 2004; Casari and Plott, 2003). More studies are needed to explore the behavioral foundations of punishment through legal systems.

Our third conclusion concerns the motivations that drive agents to punish. Two classes of motivations were considered, instrumental and expressive, and we find a considerable amount of evidence that supports the expressive punishment motivation. For an instrumental punisher the essential issue is the total punishment that an agent receives. She has no objections to others doing the "dirty job" of punishing; she actually prefers it because it saves her the punishment cost. For an expressive punisher, instead, it is the *personal* action of punishing that brings utility, regardless of others decision to punish. One could interpret it as an emotional response to a norm violation that involves no strategic considerations at all.

Without the expressive motivation to punish, it is difficult to explain the evidence of multiplicity of requests to punish the same action when the punishment choice is simultaneous (Result 4), and the evidence of its persistence when the punishment choice is sequential even

though coordination should be easier (Result 6). Moreover, early movers in the sequential design do not free ride on the punishment costs by letting later movers enforce the social norm (Result 7). These results are in line with the meta-analysis of Falk et al. (2001), where they find little evidence of strategic motifs behind punishment. Although more evidence is needed before reaching a firm conclusion on which class of models is the most accurate, this study suggests that subjects do not perceive punishment as a second-order public good and that when it comes to other-regarding attitudes, emotions seem to alter the ability of people to behave strategically.

References

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). "The Carrot or the Stick: Rewards, Punishment and Cooperation," *American Economic Review,* 93, 3, 893-902.

Bergstrom, T., Blume, L., and Varian, H. (1986). "On the Private Provision of Public Goods," *Journal of Public Economics*, 29, 25-49.

Bochet, O., Page, T., and Putterman, L. (2002). "Communication and Punishment in Voluntary Contribution Experiments," Brown University, Department of Economics, *Working Papers no. 2002-29.*

Bowles, S., Carpenter, J., and Gintis, H. (2001). "Mutual monitoring in teams: The effects of residual claimancy and reciprocity," mimeo.

Carpenter, J. (2002). "The Demand for Punishment," *Working Paper 0243*, Middlebury College, Department of Economics

Casari, M. and Plott, C.R. (2003). "Decentralized Management of Common Property Resources: Experiments with Centuries-Old Institutions," *Journal of Economic Behavior and Organization*, 51, 2, 217-247.

de Quervain, D. J.-F.., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). "The neural basis of altruistic punishment," *Science*, 305, 1254-1258.

Duffy, J., and Ochs, J. (2004). "Cooperative Behavior and the Frequency of Social Interaction," *working paper*, Department of Economics, University of Pittsburg.

Falk, A., Fehr, E., and Fischbacher, U. (2001). "Driving Forces of Informal Sanctions, *Working Paper no. 59*, University of Zurich.

Falkinger, J., Fehr, E., Gächter, S., and Winter-Ebmer, R. (2000). "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence, " *American Economic Review*, 90, 247-264.

Fehr, E. and Fischbacher, U. (2004). "Social Norms and Human Cooperation," *TRENDS in Cognitive Science*, 8, 4, 185-190.

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

Fehr, E. and Gaechter, S. (2002): "Altruistic Punishment in Humans", *Nature*, Vol.415, 137-140.

Fischbacher, U. (1998). "z-Tree: Zurich Toolbox for Readymade Economic Experiments. Instructions for Experimenters." Mimeo, University of Zurich.

Francis, H. (1985) "The Law, Oral Tradition and the Mining Community," *Journal of Law and Society*, Winter, 12, 3, 267–71.

Girard, R. (1977). *Violence and the Sacred*, Johns Hopkins University Press.

Holmstrom, B. (1982). "Moral hazard in teams," *Bell Journal of Economics*, 13, 324-40.

Horne, C. (2001). "Sociological perspectives on the emergence of norms," In: Social Norms (Hechter, M. and Opp, K.D., eds), 3-34, Russell Sage Foundation.

Kosfeld, M., and Riedl, A. (2004) The Design of (De)centralized Punishment Institutions for Sustaining Cooperation, *Tinbergen Institute Discussion Paper*, TI 2004-025/1

Lowie, R. (1970). *Primitive Society*, New York, W W Norton & Co.

Masclet, D., Noussair, C., Tucker, S., and Villeval, M.-C. (2003). "Monetary and Nonmonetary Punishment in the Voluntary Contribution Mechanism," *American Economic Review*, 93, 1, 366-380.

Masclet, D., Denant-Boemont, Noussair, C. (2004). Public Good Game with Sanctions and Metanorms, *working paper*.

Nikiforakis, N. S. (2004). "Punishment and Counter-punishment in Public Goods Games: Can we still govern ourselves?," March, *Working paper*, University of London, Royal Holloway, Department of Economics.

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

Ostrom, E. (1996). *Governing the commons : the evolution of institutions for collective action*, Cambridge, Cambridge University Press.

Putterman, L. and Anderson, C. M. (2003). "Do Non-strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism," Brown University, Department of Economics, *Working Papers no. 2003-15*.

Sefton, M., Shupp, R., and Walker, J.. (2002). "The effect of rewards and sanctions in provision of public goods," *CEDEX Working paper no. 2002-2.*

Stella, G. M. (1996). *"Schei". Dal boom alla rivolta: il mitico Nordest*, Milano, Baldini & Castoldi.

Varian, H. (1994), "Sequential Contributions to Public Goods," *Journal of Public Economics*, 53, 165-186.

Appendix A: Predictions

We outline predictions for the canonical model and for two classes of models, expressive and instrumental under the assumption of a quasi-linear utility function (4).[21]

A. *Canonical Model*

The game is Γ =(n, ({0,1,..,20}x{0,1,2,…}$_i$)$_{i\in n}$, ($u_i = \pi_i^1 - c(p^j_i)$)$_{i\in n}$). In all three treatment conditions the unique dominant strategy equilibrium is that all subjects will contribute nothing to the public good and will not punish in all periods. In fact, choosing p^j_i>0 is a monetary cost that does not generate any monetary benefit in a one-shot interaction. This is most transparent in the treatment without punishment. This condition consists of a sequence of ten (almost pure) one-shot games. In each one-shot game the agents' dominant strategy is to free ride fully. In the treatments with punishment the situation is slightly more complicated because each one-shot game now consists of two stages. It is clear that a rational money maximizer will never punish at the second stage because this is costly for the agent. Since money maximizers will recognize that nobody will punish at the second stage, the existence of the punishment stage does not change the behavioral incentives at the first stage relative to the treatment without punishment. As a consequence, everybody will choose $g_i = 0$ at stage one (Fehr and Gaechter, 2000).

B. Model of Expressive Punishment

The strength of agent *i*'s attitude toward the punishment of agent *j* is measured by the level of expressive punishment $\hat{p}_i^j$:

$$\hat{p}_i^j = \underset{p_i^j \in \{0,1,2,...\}}{\arg\max} \left\{ u_i\left(\pi_i^1, p_i^j\right) \right\} \qquad (7)$$

When $\hat{p}_i^j$ =0 agent *i* behaves as a money maximizer (5). Agent *i* wants to punish agent *j* more than agent *k* does if and only if $\hat{p}_i^j > \hat{p}_k^j$.[22]

For the expressive model, and later for the instrumental model, we present predictions only for the punishment decision and under the assumption that contribution decisions have already been taken. Except that for some examples of v_i, general equilibrium predictions for the contribution *and* punishment game have not been worked out. The punishment game when all agents are expressive punishers is Γ =(n, ({0,1,2,…}$_i$)$_{i\in n}$,

($u_i = \pi_i^1 - c(p^j_i) - v_i\ (p^j_i)$)$_{i\in n}$) and the equilibrium outcome has the following features for the three treatments:

[21] We assume an affine utility function. Notice that for simplicity there is no other-regardness attitude in the first stage.

[22] This definition, as all the discussion in this Section, is done in reference to a generic target agent *j* and not to the whole group. For this reason, we will sometimes drop the *j* subscript from expressions without fear of confusions. Notice that the number of points of punishment must be a non-negative integer. Regarding the v_i specification, compacting others' punishments into a single variable, p^j_{-i}, excludes preference types that considers how the punishment is carried out, i.e. some agents may be willing to punish only if others also do it, which may be relevant especially in the Sequential treatment.

Predictions for the Baseline treatment:

BAS 1E: The total level of punishment is given by the *sum* of each agent's expressive punishment, $\boldsymbol{P^*}=\sum_{i=1}^{n}\hat{p}_i^j$.

BAS 2E: *All* agents *i* with $\hat{p}_i^j>0$ participate in the punishment.

BAS 3E: There trivially exists a *unique* Nash equilibrium (a dominant equilibrium).

Predictions for the Consensual treatment:

CON 1E: There is *equal or less* punishment under Consensual than under Baseline treatment. Punishment is strictly less for preference profiles where only one agent *i* has $\hat{p}_i^j>0$.

CON 2E: There trivially exists a *unique* Nash equilibrium (a dominant equilibrium).

Predictions for the Sequential treatment:

SEQ 1E: The overall punishment level ***P**** is *equal* to the punishment assigned under Baseline treatment.

SEQ 2E: *All* agents *i* with $\hat{p}_i^j>0$ participate in the punishment. There are no differences in punishment patterns between the Baseline and Sequential treatments.

SEQ 3E: First and last movers have on average the same burden of punishing.

Proofs are trivial as there are no strategic considerations at all in punishment choices because agents' best replies are independent of p^j_{-i} .

C. Model of Instrumental Punishment

This framework borrows many ideas and details from Varian (1994). The present study focuses on preferences for punishment while Varian (1994) discusses preferences for public good contributions. When $\bar{p}_i^j>0$, agent i will punish agent *j* if nobody else does, meaning that she is willing to enforce, at least partially, a social norm at a private monetary cost. She will not if another agent *k*, who likes to punish more than agent *i*, $\bar{p}_k^j>\bar{p}_i^j$, has already punished.[23]

The equilibrium strategies in the three experimental designs for the punishment game when all agents are instrumental punishers $\Gamma=(n, (\{0,1,2,\ldots\}_i)_{i\in n}, (u_i=\pi_i^1-c(p^j_i)-v_i(p^j_i, p^j_{-i}))_{i\in n})$ are now computed. We assume that agents' $\bar{p}_i^j$ are common knowledge. Throughout the analysis we illustrate the equilibrium strategy for *n*=3, 4, 5 but the results can be easily extended to an arbitrarily large *n*.

First design: Baseline punishment. When all agents simultaneously announce punishment requests, from the first order condition, $\partial u_i/p_i=0$, one can compute an agent's best reply $B_i(p_{-i})$. Sometime we will drop the superscript *j* for

[23] Notice that $\bar{p}_i^j=\hat{p}_i^j$.

the target agent to simplify notation. From $v_i'(p_i + p_{-i})=c'$, we obtain $B_i(p_{-i})= \overline{p}_i - p_{-i}$. When $p_i>0$, we have $B_i(p_{-i})=\max\{0, \overline{p}_i - p_{-i}\}$.

We assume that agents' standalone punishment levels are common knowledge. The Nash equilibrium of the punishment game depends on the set of agents with the maximum punishing attitude, $M=\{i\neq n: \max_{k\neq n}\{ \overline{p}_k \} = \overline{p}_i \}$. There are two possible cases. When there is a single agent in the set, $|M|=1$, the equilibrium strategy is unique and is a corner solution where the agent i who most like punishing punishes $\overline{p}_i$ and all the others $j\neq i$ punish zero. When $|M|>1$, there are multiple Nash equilibria that all yield a unique aggregate amount of punishment.

$$\text{Nash equilibrium: } \{i\in M,\ p_i^k*: \Sigma_{i\in M}\, p_i^k* = \overline{p}_i^k \text{ and for } j\notin M,\ p_j^k*=0\} \qquad (9)$$

For example when $\overline{p}_1=6$, $\overline{p}_2=4$, $\overline{p}_3=1$, $\overline{p}_4=0$, in equilibrium agent n receives 6 punishment points from agent 1 and zero from the other agents. In summary, the equilibrium outcome has the following features:

BAS 1I: The equilibrium level of punishment is always uniquely set at $\boldsymbol{P}^*= \overline{p}_i$, where agent i has the *maximum standalone punishment level*

BAS 2I: The burden of punishing falls on the subject(s) with the maximum standalone punishment level, $p_i^*\geq 0$ for $i\in M$ and $p_j^*=0$ for $j\notin M$.

BAS 3I:There exists a *unique* Nash equilibrium when there is just one agent with a strictly maximum standalone level, $|M|=1$, and there exist *multiple* Nash equilibria otherwise, $|M|>1$.

Example. Consider a game with $n=3$ agents where the punishment attitude is instrumental and modeled by the anonymous function $v_i\,(p^j_i, p^j_{-i}) = \alpha_i \ln (p^j_i + p^j_{-i})$, where $\alpha_i = f(g_1,.., g_n)$. Assume that agent 3 never punishes, $\alpha_3=0$ while agents 1 and 2 punish similarly and only when contributions are below a common threshold: for $i=1,2$ $\alpha_i = \{$for $j=1,2,3$, $j\neq i$, $\alpha_i=\gamma>0$ when $g_j < g'$; $\alpha_i=0$, otherwise$\}$. In the Nash equilibrium of the *punishment game* the amount of punishment received by agent 3 is $p^3_1* + p^3_2*=\gamma$. This is derived from the FOC, $\partial v_i/\partial p^j_i = \partial c/\partial p^j_i$: $\gamma/(p^j_i + p^j_{-i})=1$ hence the best reply for $i=1,2$ is $p^j_i*=\{\gamma - p^j_{-i}$, if $g_j < g'$; and 0 otherwise$\}$. One can work out the general equilibrium of the *contribution and punishment game* for agent 3 by plugging p^j_i* into $u_3 = \pi^1_i - e(p^3_1* + p^3_2*)$ and then finding an optimal contribution level, $g_3*=\text{argmax}\{u_3\}$. A positive contribution is optimal, $u_3(g_3=0) < u_3(g_3=g')$ when $e(\gamma) > (1-\lambda)\, g'$. Given $e(\gamma)=3\gamma$ and $\lambda=0.4$, there is an equilibrium with partial contribution at $g_i*=10$ when $\gamma>2$, $g'=10$ and an equilibrium with full contribution at $g_i*=20$ when $\gamma>4$, $g'=20$. The same logic applies to agents 1 and 2. In a general equilibrium with positive contribution, $((g_1*, g_2*, g_3*), (p^1*, p^2*, p^3*))=((g',g',g'), (0,0,0))$.

Besides this threshold example, the functional form for $v_i\,(p^j_i, p^j_{-i}) = \alpha_i \ln (p^j_i + p^j_{-i})$ can accommodate a variety of attitudes toward punishment. For instance a preference for equal shares, $\alpha_i = \max\{ 0, \gamma_i\,(1/n - g_j/G) \}$ with $\gamma_i\geq 1$, or an aversion toward disadvantageous inequality, $\alpha_i = \max\{ 0, \gamma_i\,(g_i - g_j) \}$ with $\gamma_i\geq 0$. One could also model anti-social behavior, such as an aversion toward cooperators, $\alpha_i = \max\{ 0, \gamma_i\,(g_j - g_i) \}$ with $\gamma_i\geq 0$, or a Falkinger-type

punishment $\alpha_i = \max\{ 0, \gamma_i (g_j - G_{-j}/n-1)\}$ with $\gamma_i>0$ (Falkinger et al., 2000). We remain agnostic on what is the appropriate specification but stress the flexibility of the model chosen.

Second design: Consensual punishment. While the game comprises *n* players, for the Consensual design it is more useful to reason in terms of the number of agents $s \leq n$ with a strictly positive standalone punishment value, s=|{i: $\overline{p}_i^k$ >0}|. In this respect, we can classify group preference profiles for *n*=5 into five exhaustive classes, N_s corresponding to s=0, 1, 2, 3, 4. When the profile is of class N_s, we consider a punishment game among just s agents because agents with $\overline{p}_i$ =0 have no effect on the equilibrium. The notation N_0 identifies situations where $\overline{p}_i$ =0 $\forall i$, N_1 situations where $\exists! i : \overline{p}_i > 0$, N_2 situations where $\exists! i, j : \overline{p}_i, \overline{p}_j > 0$, and so on. The Nash equilibrium of the punishment game in the Consensual treatment is:

$$\{\text{for } s=0,1,\ p_i^{k}*=0\ \forall i\in n;$$

$$\text{for } s>1 \text{ and } |M|=1,\ \exists i\in M:\ p_i^{k}*= \overline{p}_i^k -1,\ \exists j\in n:\ p_j^{k}*=1, \text{ and for } \forall t\in n: t\neq i,j,\ p_t^{k}*=0;$$

$$\text{for } s>1 \text{ and } |M|>1,\ i\in M,\ p_i^{k}*:\ \Sigma_{i\in M}\, p_i^{k}*=\max\{2, \overline{p}_i^k\} \text{ and for } j\notin M,\ p_j^{k}*=0\} \qquad (10)$$

While the s=0,1 case is trivial, few comments will be made on the two cases with s>1. When the s agents exhibit identical attitudes toward punishment, $\overline{p}_i = \overline{p} > 0\ \ \forall i$, there are multiple Nash equilibria that all yield a unique amount of punishment $\boldsymbol{P}^*= \overline{p}$. An agent's best reply is $B_j(p_{-j}) =\max\{0, \overline{p}_j - p_{-j}\}$. This result is analogous to the one under Baseline rule[24] with the difference that now the equilibrium set for the Consensual rule is a subset of the equilibrium set for the Baseline rule. This latter point follows from the consideration that, in order for a sanction to be effective, at least two agents must request a positive amount of punishment.

When the s agents exhibit diverse attitudes toward punishment, the Nash equilibrium strategy must as well involve strictly positive punishment from at least two agents. We now describe the Nash equilibrium set for the five classes of preference profiles. The outcome under N_0 is trivial. Under N_1 there is never punishment with a Consensual rule. Under N_2 the best reply is $B_i(p_{-i})=(\overline{p}_i - p_{-i})$ for the agent with the highest standalone punishment $i = \underset{k}{\text{argmax}}\{\overline{p}_i\}$ and $B_j(p_{-j})=1$ for the other. The unique Nash equilibrium is then $(\overline{p}_i - 1, 1)$. The second agent *j* never prefers to punish more than one point because agent *n* is already punished more than agent *j*'s standalone value, $\overline{p}_i - 1 + p_j$

[24] Due to the integer nature of p_i, there is an exception in the equilibrium when $\overline{p}$ =1. In particular, two individuals give one punishment point each and hence agent *k* receives two points of punishment.

> $\bar{p}_j$. Moreover, agent *j* always prefers one punishment point to zero because $u_j(\pi_j^1, 1, \bar{p}_i - 1) > u_j(\pi_j^1, 0, 0)$.[25] On the other hand, agent *i*'s punishment request will have an effect only when the other agent joins him in the punishment effort. As the second agent *j* never puts more than one point of punishment, agent *i*'s equilibrium strategy is to request $\bar{p}_i - 1$ points of punishment.

Under N_3 and N_4 there are multiple Nash equilibria. All Nash equilibria are characterized by a constant punishment level of agent *n*, $\boldsymbol{P^*} = \bar{p}_i$, where $i = \underset{k}{\operatorname{argmax}}\{\bar{p}_i\}$ and $p_j^* > 0$ for two and only two agents. The best reply for the agent or the agents *i*: $\bar{p}_i = \max_j\{\bar{p}_j\}$ is $B_i(p_{-i}) = (\bar{p}_i - p_{-i})$ and for just one of the other agents *i*: $\bar{p}_i < \max_j\{\bar{p}_j\}$ is $B_j(p_{-j}) = 1$. In comparison with N_2, there is the additional problem of coordinating among the s agents. For example when $\bar{p}_1 = 6$, $\bar{p}_2 = 4$, $\bar{p}_3 = 1$, in equilibrium agent *n* receives 6 punishment points; 5 points from agent 1 and one point from either agent 2 or agent 3. When there is miscoordination more than two agents may end up punishing.

In conclusion, there are three differences in outcome between the Baseline and the Consensual rules of punishment.

CON 1I: Under Consensual rule there is *equal or lower* overall punishment level than under Baseline rule.

CON 2I: While the burden of punishment can fall on just one agent under Baseline rule, it is always *spread between two (or more) agents* under Consensual rule.

CON 3I: While with identical preferences for punishment the set of Nash equilibria is smaller under the Consensual rule, with diverse preferences it is larger.

Third design: Sequential punishment. We analyze the equilibrium strategies under the Sequential punishment separately for *n*=5 for each of the five classes of preference profiles. We assume that the identity of each agent at each step is known. The outcome under N_0 is trivial. Under N_1 the outcome is the same as under Baseline rule, *i*.e. $P = \bar{p}_1$. Under N_2 consider the order of moves where agent 1 moves first and agent 2 second. This analysis has been adapted and expanded from the contribution of Varian (1994). The best reply of agent 1 can be derived from his indirect utility function:

$$z_1(\pi_1^1, p_1) = \pi_1^1 - p_1 + v_1(p_1 + B_2(p_1)) \qquad (11)$$

$$z_1(\pi_1^1, p_1) = \pi_1^1 - p_1 + v_1(p_1 + \max\{(\bar{p}_2 - p_1), 0\}) \qquad (12)$$

Let us consider two possible cases. When preferences are identical $\bar{p}_i = \bar{p} \ \forall i$, agent 1 (the first mover) assigns zero points of punishment to agent k and agent 2 (the second mover) chooses $\bar{p}$. The level of punishment **P** is the same as under Baseline rule. The difference is that now there is a unique Nash equilibrium where the second mover

[25] This condition is equivalent to $v_j(\bar{p}_i) - v_j(0) > 1$, which can be rewritten as $v_j(\bar{p}_i) - v_j(\bar{p}_j) + v_j(\bar{p}_j) - v_j(0) > 1$. The condition is satisfied since $v_j(\bar{p}_i) - v_j(\bar{p}_j) > 0$ because v_j is increasing and $\bar{p}_i > \bar{p}_j$, and $v_j(\bar{p}_j) - v_j(0) > 1$ because $\bar{p}_j > 0$.

always bears all the cost of punishing. Loosely speaking, the order of moves solves the coordination problem among agents, which is present under Baseline rule.

When preferences are diverse, the outcome depends on the relative preferences for punishment of the two agents, 1 and 2, and assume that agent 1 announces her punishment request first. Consider the case where agent 2 likes to punish the most, $\bar{p}_2 > \bar{p}_1$. In this case the optimal strategy for agent 1 is to choose zero punishment. Now consider the case where agent 2 likes to punish the least. Agent 1's optimal strategy is then to punish for the whole amount $\bar{p}_1$ only if he likes to punish *much* more than agent 2 and to punish zero otherwise. More formally, the condition

$$z_1(\pi_1^1, 0) < z_1(\pi_1^1, \bar{p}_1) \qquad (13)$$

reduces to $\Delta_1 > \bar{p}_1$ where $\Delta_1 = v_1(\bar{p}_1) - v_1(\bar{p}_2)$. The intuition behind this strategy is that agent 1 chooses between not punishing, hence getting the preferred punishment level of agent 2, and fully paying for his preferred level of punishment, which is higher. He will punish if the additional utility of the higher punishment is worth the cost. On the other hand, when preferences are similar, $\Delta_1 < \bar{p}_1$, the optimal strategy is zero punishment as in the case of identical preferences.

Consider an example with the following utility function:

$$u_i(\pi_i^1, p_i, p_{-i}) = \pi_i^1 - p_i + \alpha_i \ln(p_i + p_{-i}), \text{ with } \alpha_i > 0 \qquad (14)$$

The preference for punishment v_i is increasing and concave, $v' = \alpha_i/(p_i + p_{-i}) > 0$ and $v'' = -\alpha_i/(p_i + p_{-i})^2 < 0$. The standalone punishment level is $\bar{p}_i = \alpha_i$. The best reply function under Baseline rule is $B_i(p_{-i}) = \max\{0, \alpha_i - p_{-i}\}$.

Under Sequential rule the indirect utility function of the first mover when there are two agents is $z_1(\pi_1^1, p_1) = \pi_1^1 - p_1 + v_1(p_1 + B_2(p_1)) = \pi_1^1 - p_1 + \alpha_1 \ln(p_1 + \max\{0, \alpha_2 - p_1\})$. Agent 1 as first mover does not punish when $\alpha_1 = 4$, $\alpha_2 = 2$ but does punish when $\alpha_1 = 6$, $\alpha_2 = 2$. In general agent 1 punishes if $\ln(\alpha_1/\alpha_2) > 1$.

The equilibrium strategy under Sequential rule is characterized by the following:

SEQ 1I: Under Sequential rule, the overall punishment level is on average *equal or lower* than under Baseline rule

SEQ 2I: The punishment is carried out by *only one agent*

SEQ 3I: On average the burden of punishing falls disproportionately on the punisher who moves *last*

SEQ 4I: There exists a unique Nash equilibrium

SEQ 5I: Predictions 1I-3I hold when agents know their own order of move but ignore other agents' order of move.

Appendix B: Instructions

The following instructions were originally written in Italian. We document the instructions we used in the first part of the experiment, which were common to all treatments, and present the second part instructions for the Consensual treatment. The instructions for the Baseline and Sequential treatments were adapted accordingly. They are available upon request.

You are now taking part in an economic experiment on decision-making. If you read the following instructions carefully you can, depending on your decision, earn a considerable amount of money.

During the experiment we shall not speak of Euros but rather of Tokens. During the experiment your entire earning will be calculated in Tokens. At the end of the experiment the total amount of tokens you have earned will be converted in Euros at the following rate

1 Token = 2 cent. Euro

At the end of the experiment your earnings will be privately paid in cash. To the amount on your screen you must add 5 Euro as a lump sum participation fee.

During the experiment you will not be asked to reveal your identity and your name will not be associated with the decisions you are going to take. Moreover, you are not allowed to talk or otherwise communicate with the other participants during the experiment.

This experiment is divided into two different parts. The following instructions are related to the first stage. The first stage consists of 10 periods.

Your earnings depend on your decision and on other four participants' decisions. The experiment participants will be randomly re–matched after each period and therefore it is highly likely that in each period you will interact with different people. You do not know the identity of the people with whom you interact.

At the beginning of each period each participant receives 20 points. You task is deciding how you would like to use these tokens. The other participants will face the same scenario. You have to decide how many points of the 20 available you want to contribute to a project. For each point that you keep for yourself you earn an income of one Token. The points you have contributed to the project plus the points that all the other four persons have contributed are converted in a double quantity of Tokens, which will be evenly divided among these five persons. Therefore, after being doubled, you will receive one fifth of the Tokens contributed to the project. To sum up your income consists of two parts:

Your income this period = direct income + income from the project

= (20 – your contribution to the project) + $^1/_5$ x ((sum of yours' and other four people's contribution to the project) x 2)

Each of the four persons will receive from the project the same amount that you will. For example, suppose the sum of the contributions of the five persons is overall 60 points. In this case each person receives from the project 60x2/5=24 Tokens. Instead, if the total contribution to the project is 10 points, each of the five persons receives an

income from the project of 10x2/5=4 Tokens. The following table gives you some examples of income from the project:

Sum of the points contributed	0	10	20	30	40	50	60	70	80	90	100
Income from the project for each of the 5 persons	0	4	8	12	16	20	24	28	32	36	40

For each point that you keep for yourself you earn an income of 1 Token. Supposing you contributed this point to the project instead, then the total contribution to the project would rise by one point. Your income from the project would rise by 1x2/5=0,4 Tokens. Your contribution to the project would also raise the incomes of other persons. More precisely, the other four persons will earn an additional 0.4 Tokens each, so that the overall income increase for you and the others would be of 2 Tokens.

After everybody has completed his or her decision, you shall see your period income on the computer screen. Moreover, there will be presented the points contributed to the project by each one of the four persons that could contribute with you as well as their period income. The identity of these other people will change randomly each period.

This procedure will be repeated 10 periods.

Are there any questions? If you have questions during the experiment we kindly ask you to raise your hand and somebody will assist you in private.

These are the instructions for the second and last part of the experiment. As before, the experiment consists of ten periods and in each period you have to make a decision about how many of the 20 tokens available to you.

Different than from before, each period is now composed of two phases, the first phase is identical to the procedure already described, while in the second phase you may choose to reduce the earnings of other people that have profited from the same project.

<u>In the first phase</u> of a period, you have to make the same type of decision as the one in the sequence already described before, and that will here be repeated.

At the beginning of each period each participant receives 20 points. You task is deciding how you would like to use these tokens. The other participants will face the same scenario. You have to decide how many points of the 20 available you want to contribute to a project. For each point that you keep for yourself you earn an income of one Token. The points you have contributed to the project plus the points that all the other four persons have contributed are converted in a double quantity of Tokens, which will be evenly divided among these five persons. Therefore, after being doubled, you will receive one fifth of the Tokens contributed to the project. To sum up your income consists of two parts:

Your income for phase one = direct income + income from the project

= (20 – your contribution + $^1/_5$ x ((sum of yours' and other four people's

to the project) contribution to the project) x 2)

After everybody has completed their decision, you shall see on the computer screen the points contributed to the fund by every one of the four persons that could contribute with you as well as their period income. Your decision and result will be shown in the first column. The identity of these other people will change randomly each period.

In the second phase of a period you can reduce or leave equal the income of each of the four persons that have profited from the same project. Conversely, the other persons can lower your earnings as well.
Your decision is about distributing points to the other four persons. There is no way for you to know the identity of the other persons because they have been randomly selected every period among all participants. You have to choose a number of points for each person and you know only his/her contribution decision in the first phase of the period. If you do not want to change the earnings of a person choose 0. If you want to reduce the earnings of a person, you can distribute a number of points from 0 to 7. For each point distributed, the income in that particular person will be reduced by 3 (THREE) tokens. For the person distributing the point, each point costs 1 token. Your overall cost is equal to the sum of the points that you have distributed to each one of the other four persons. Your maximum cost for distributed points is then 28 tokens (7 tokens times 4 persons). Your cost is zero if you do not distribute points to anybody.

As it will now be explained, a request to distribute points is not always carried out. For each person, there are two cases.
When ONLY YOU have requested to distribute points to a given person, your decision has no effect. In particular, there is no reduction in his/her income and no payment on your side for your request. In the opposite case, when BOTH YOU AND OTHERS have requested to distribute points to that same person, then your decision to distribute points is carried out. Requests by others to distribute points to that person will also be carried out. In other words, there have to be at least two requests to distribute one or more points to the same person in order to carry out a reduction of his/her income. It does not matter that the two requests are for distinct amounts.

EXAMPLES. If you distribute 0 points to a person you do not change his/her income. Suppose you request to distribute 6 point to a person. Under some conditions this request does not have any effect, while under other conditions you reduce his/her income by 18 tokens (6x3). More precisely, if all the others distribute 0 points, your request will be ignored. This result will be signaled on the screen at the end of each period by the note “Points distributed? NO” in the column corresponding to the concerned person. On the contrary, if at the same time somebody else has distributed for instance 2 points to the same person, your request is carried out (an 18–token reduction) and you will be charged the fee of 6 tokens. In addition, the request of the other person will be carried out. The cumulated effect of the two requests is an overall income reduction of (6+2) x 3 = 24 tokens. This result is marked on the screen at the end of the period by the note “YES” in the column corresponding to the concern person.

Your total income at the end of the period will be:

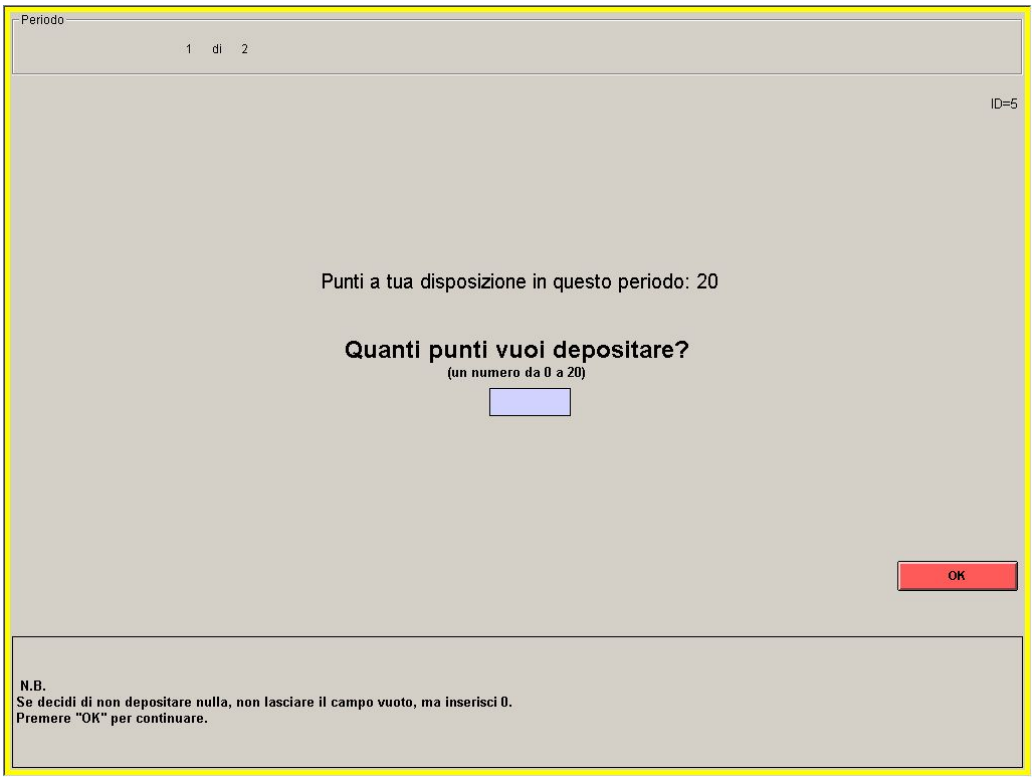
Your period income = phase one income – income reduction – cost to distribute points

= phase one income – (sum of received points)x 3 – (total points distributed).

After everybody has completed their decision, you shall see on the computer screen the results for phase two. For each person you will learn the cumulative income reduction due to the points distributed. Individual requests to distribute points will remain confidential in order to preserve the anonymity of decisions.

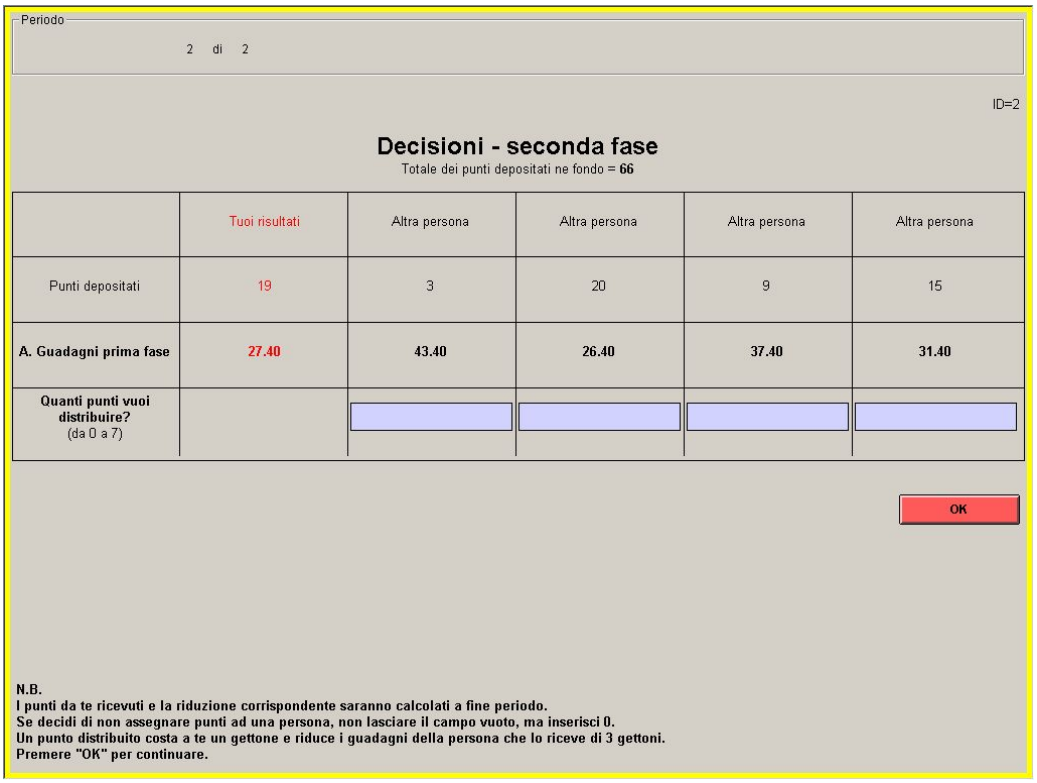
Are there any questions?

Screen shot 1: Project contribution decision, all treatments



Screen shot 2: First phase results, all treatments

Screen shot 3: Input screen for second phase, all treatments



Screen shot 4: End of period results, Consensual treatment

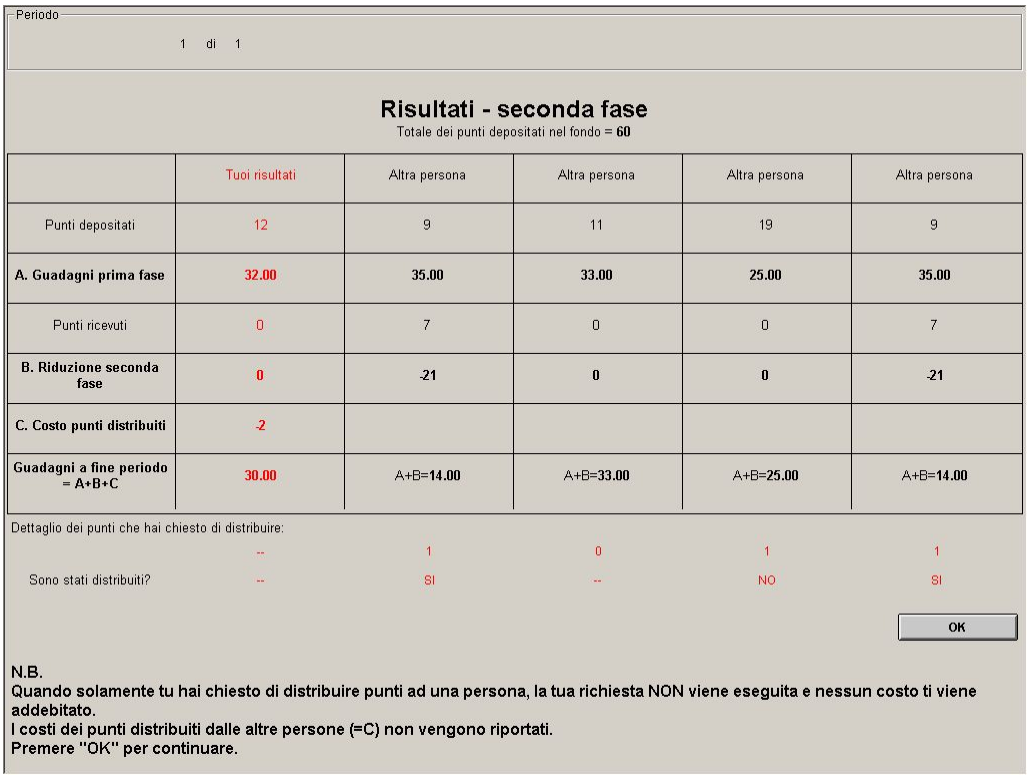


Table 1 – Treatment Conditions

	Baseline	*Consensual*	*Sequential*
Decisions to contribute are Simultaneous	Yes	Yes	Yes
Decisions to punish are Simultaneous	Yes	Yes	No
Punishment when only one agent requested it	Yes	No	Yes
Number of sessions	4	4	4
Total number of participants	80	80	80
Periods without punishment opportunity	10	10	10
Periods with punishment opportunity	10	10	10

Table 2: Summary of individual contribution across experimental sessions

Treatment	*Baseline*				*Consensual*				*Sequential*			
Session date	3/27	10/15	10/17	10/23	5/22	9/16	10/16	10/21	4/14	10/16	10/20	10/22
avg (sd)												
No	3.54	4.40	2.57	3.34	2.56	5.07	3.91	2.27	2.80	4.00	3.07	2.19
punishment	(5.58)	(6.08)	(4.34)	(5.56)	(4.13)	(6.26)	(5.49)	(4.09)	(4.85)	(5.66)	(4.92)	(3.62)
With	7.74	5.14	2.62	2.36	14.11	11.26	5.89	2.57	4.18	5.53	2.42	5.41
punishment	(5.42)	(5.25)	(3.12)	(2.49)	(6.54)	(6.59)	(4.51)	(2.93)	(5.68)	(4.36)	(4.57)	(4.87)

Table 3: Punishment rates by individual level of contribution (Sub-sample)

Individual contribution level	Baseline	Sequential	Consensual			
	Avg. points *(no. obs.)*	Avg. points *(no. obs.)*	Assigned (1)	Requested (2)	Difference (2) – (1)	% Censored [(2) – (1)] /(2)
0	5.064 *(124)*	4.497 *(177)*	4.221 *(68)*	4.500	0.279	6.2%
(0, 5]	1.663 *(261)*	2.345 *(171)*	1.006 *(159)*	1.434	0.428	29.8%
(5, 10]	0.855 *(83)*	1.464 *(153)*	0.446 *(83)*	1.060	0.614	57.9%
(10, 20]	0.778 *(27)*	1.424 *(59)*	0.200 *(30)*	0.933	0.733	78.6%
Total	2.331 *(495)*	2.687 *(560)*	1.441 *(340)*	1.912	0.471	24.6%

Notes: To partially control for the uneven distribution of contribution levels across treatments, the table includes only the sub-sample of experimental data where *group* contribution was in [10, 40], which constitutes about 58% of the sample (1395/2400 obs.).

Table 4: Frequency of punishment

	Baseline	*Sequential*	*Consensual*
Contribution choices not punished	**32.1%**	**27.4%**	**66.3%**
Of which: One request to punish	-	-	26.3%
Contribution choices punished	**67.9%**	**72.6%**	**33.9%**
Of which: One request to punish	32.5%	35.8%	-
Two requests to punish	20.0%	23.1%	18.8%
Three requests to punish	12.0%	11.1%	10.4%
Four requests to punish	3.4%	2.6%	4.6%
Total	100.0%	100.0%	100.0%
(observations)	(800)	(800)	(800)

Table 5: Variance decomposition of individual requests to punish – Baseline treatment

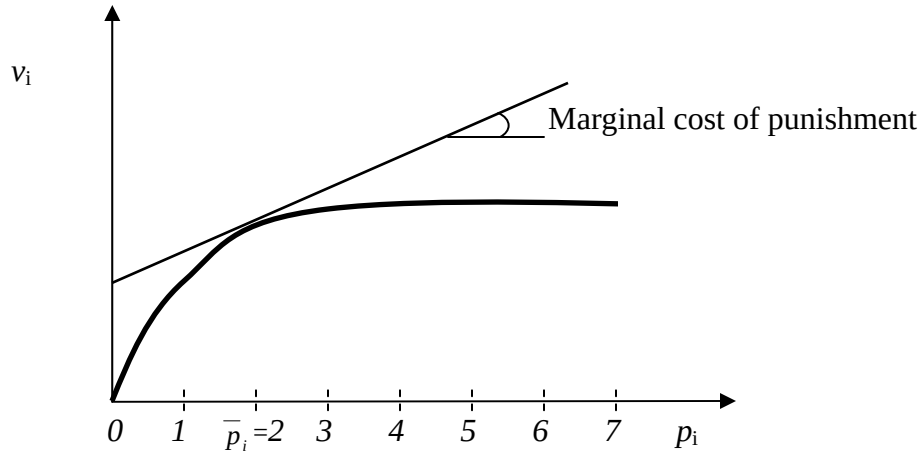
No. of requests to punish	*Total variance* $Var[p^j_i]$	*Variance across subjects* $Var_x[E[p^j_i\|j]]$	*Variance within subjects* $E_x[Var[p^j_i\|j]]$	*Obs.*
1	1.451	1.451	-	318
2	3.462	1.023	2.439	188
3	4.040	1.040	3.000	102
4	4.421	0.861	3.560	29

Table 6: Determinants of getting punished: regression results

	Baseline	*Sequential*	*Consensual*	*All treatments*
High cooperators	0.7999	0.2274	-0.8794	-0.1897
(contributions >15 tokens)	(0.8858)	(0.5818)	(0.3183)*	(0.3975)
Average contribution of	-0.0945	-0.0172	0.0569	0.0025
others in the group	(0.1612)	(0.0358)	(0.0193)*	(0.0438)
Positive deviation from	-0.0707	0.0018	-0.0006	-0.0223
average	(0.0648)	(0.0324)	(0.0134)	(0.0252)
Absolute negative deviation	0.7871	0.6295	0.5688	0.6329
from average	(0.0768)***	(0.1142)**	(0.0434)***	(0.0538)***
Constant	1.5221	-0.6236	0.1575	1.0468
	(1.6952)	(0.3496)	(0.3700)	(0.4299)**
No. of observations	800	800	800	2400
R-squared	0.39	0.41	0.57	0.45

Notes: OLS estimator clustered by session. It includes session and period dummies, not reported. Robust standard errors in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%

Figure 1: Utility from punishment



Note: The utility v_i that agent *i* enjoys from punishing agent *j* is increasing in the punishment level p_i. The standalone punishment level $\bar{p}_i$=2 is also equal to the optimal punishment level for an expressive punisher.

Figure 2: Gross contribution levels

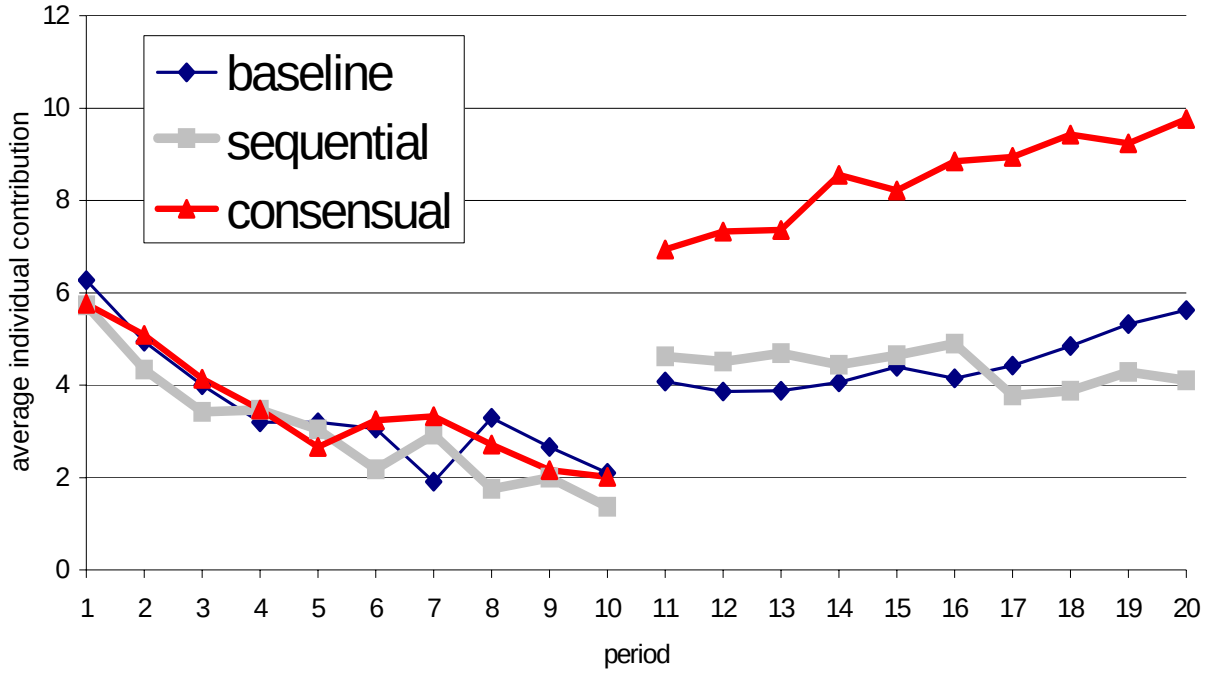
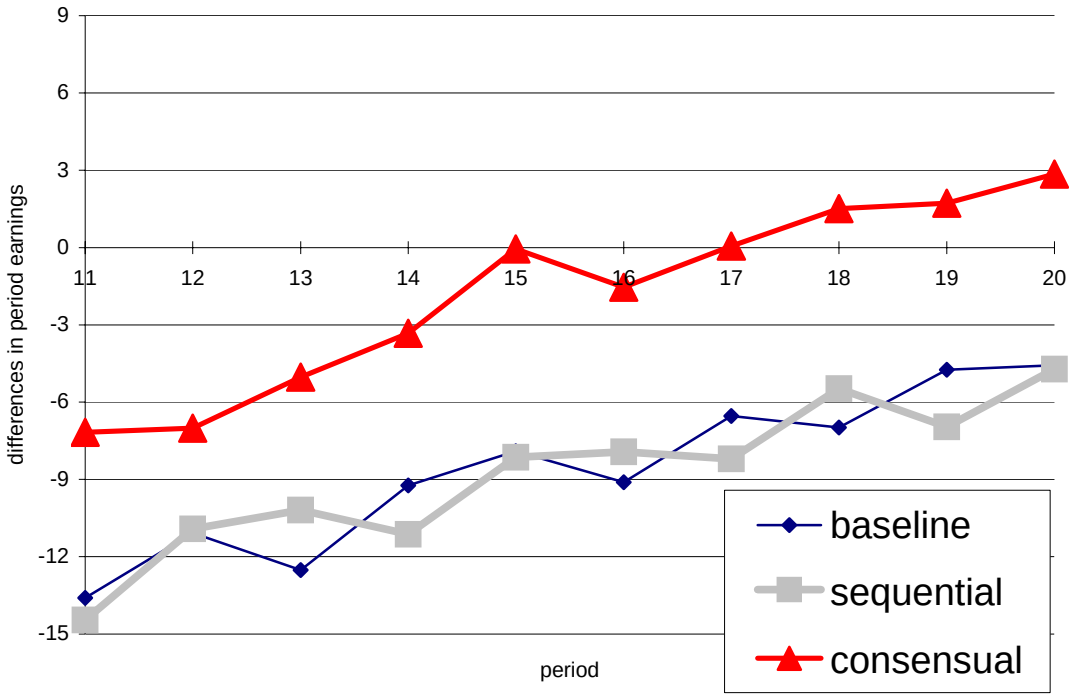


Figure 3: Net aggregate surplus over time



Note: Period by period difference between sequence with sanction and without sanction

Figure 4: Average punishment by treatment

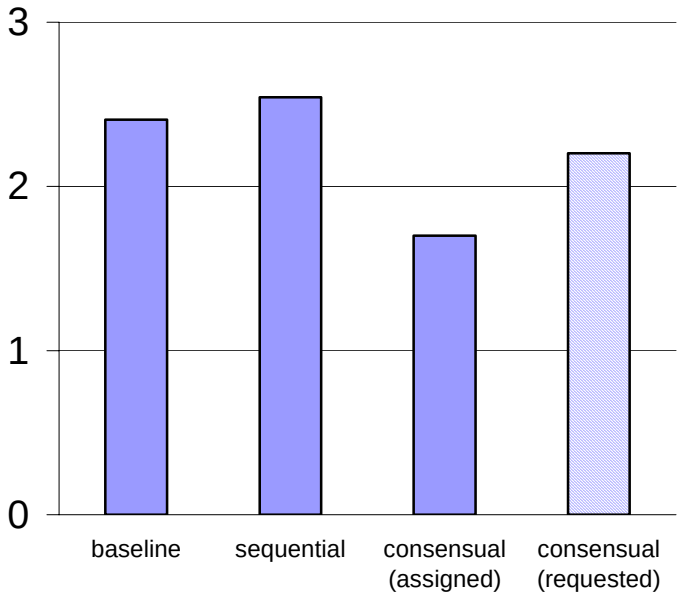


Figure 5: Empirical distribution of individual punishment requests

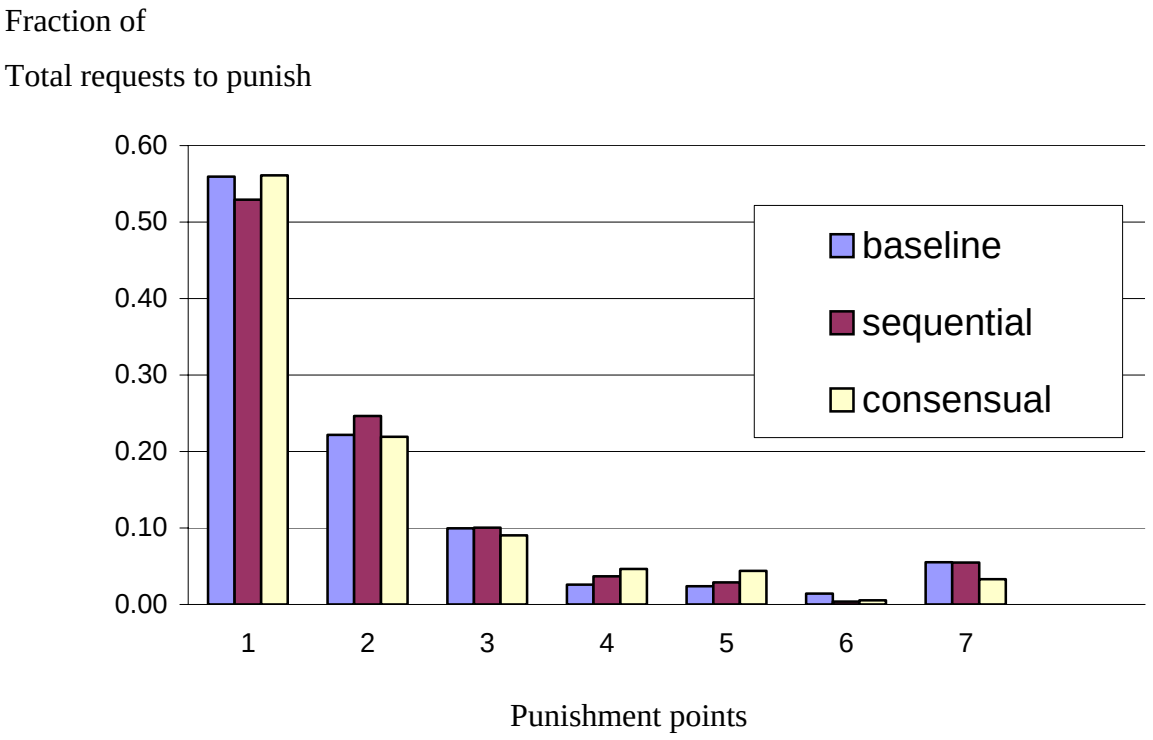


Figure 6: Sequential punishment step by step