Peer Punishment in Teams:

Expressive or Instrumental Choice?

Marco Casari and Luigi Luini *

August 27, 2009

Abstract

A key question about human societies is how social norms of cooperation are enforced. Subjects who violate norms are often targeted by their peers for punishment. In a public good game with peer punishment, the punishment of norm violators constitutes a second-order public good. In an experiment we examine whether subjects do treat punishment itself as a public good. Results do not support this view and rather suggest a hard-wired taste for punishment, which lets subjects ignore the public good characteristics of punishment.

JEL C91, C92, D23

Keywords: experiments, sanctions, public goods, strategic behavior, social norms

---

* Contact information: Marco Casari (corresponding author), Università di Bologna, Department of Economics, Piazza Scaravilli 2, 40126 Bologna, Italy, Tel. +39 051 209 8662, Fax +39 051 209 8493; 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. We thank Francesco Lomagistro for programming and help in running the sessions. This version of the manuscript has benefits from comments from Tim Cason, Nikos Nikiforakis, Steve Gjerstad, Pedro del Rey, Gabriele Camera, Samuel Bowles, seminar participants at ESA meetings in Alessandria, Italy, Purdue University, USA, University of Bologna in Forlì, University of Melbourne and Monash University, Australia, and two anonymous referees. The usual disclaimer applies.

1 Introduction

This paper sheds light on how people enforce social norms of cooperation. Experiments with social dilemmas have uncovered a robust behavioral tendency to engage in peer punishment of norm violators, which is often in contrast with the predictions derived from models of rational self-regarding agents (e.g. Ostrom et al., 1992; Andreoni et al., 2003; Guererk et al., 2003; Gächter et al., 2008; Nikiforakis and Normann, 2008); yet scholars have only a partial understanding of what motivates punishment. This paper studies the view of peer punishment as an "instrument" that subjects employ because of its final income consequences on the targeted agent. This view is an articulation of what in the literature is sometimes called punishment as second-order public good:

> "If those who free ride on the cooperation of others are punished, cooperation may pay. Yet this `solution' begs the question of who will bear the cost of punishing the free riders. Everybody in the group will be better off if free riding is deterred, but nobody has an incentive to punish the free riders. Thus, the punishment of free riders constitutes a second-order public good. The problem of second-order public goods can be solved if enough humans have a tendency for altruistic punishment, that is, if they are motivated to punish free riders even though it is costly and yields no material benefits for the punishers." (Fehr and Gächter, 2002, p.137)

The main contribution of this paper is to put forward a formal framework for this widespread view of peer punishment, spell out some of its implications, and carry out an experiment to study it. In the most common design for peer punishment experiments, multiple subjects can simultaneously punish the same target. This design is unfit to study how people enforce social norms of cooperation." In a simultaneous design the observed punishment choices may be due to a subject's taste for punishment or to strategic considerations and it is hard to disentangle the two in order to test the instrumental model. When a subject punishes instrumentally, she values

having the person targeted receiving a certain amount of punishment but does not care to do it personally, especially because it is costly. As a consequence, her choice will strategically depend on how much others do or will punish. This game will likely admit multiple equilibria. This point will be explained in detail in Section 2,

To obtain cleaner evidence for or against instrumental punishment, we designed a new experiment (Sections 3 and 4). We begin with a "one-to-one" treatment where there is no strategic consideration in punishment and employ the experimental results as a benchmark to measure the individual taste for punishment. In a "sequential" treatment we measure how the same subjects react when there are strategic considerations in punishment.

The results are reported in section 5 and, contrary to our expectation, do not support the instrumental punishment model. This evidence poses a puzzle as this simple, intuitive model does not explain the data. In search of a better model, section 6 puts forward two conjectures, reciprocity in punishment and expressive punishment.[1] Reciprocity in punishment implies that a subject will punish a norm violator only if others punish her as well. According to this view, if a subject observes someone who had the opportunity to punish a norm violator but did not do it, then also the subject will not punish the norm violator. The data do not support this conjecture. The other conjecture concerns expressive punishers who punish because they obtain satisfaction from engaging in the act of punishing. As an implication of this view, another member's punishing is not a substitute of personal punishment. If punishment is expressive, the group outcome may have tenuous relations with the rationale of legal punishment systems, which follows "punishment must fit the crime" rule. We conclude in section 7 that the most promising conjecture is expressive punishment.

[1] Expressive punishment is equivalent to the "emotional punishment" of Casari and Luini (2009).

2 When punishment is instrumental

Let us assume that agent *i* may have a taste for punishment as part of her utility function u_i. Consider three agents who have contributed to the group projects amounts g_1, g_2, and g_3, respectively. Agent *i* may want to punish agent *k* a number of points p_{ik} for reasons related to the social norms or other-regarding preferences of agent *i*; for instance because agent *k* is a free-rider. Let us allow heterogeneity in agents' taste for punishment. To measure its intensity we introduce the concept of standalone punishment: s_{ik} denotes agent *i*'s taste for punishment versus agent *k*, when nobody else punishes, $\sum_{j \neq i} p_{-jk} = 0$,

$$\boxed{\quad} \qquad (1)$$

The standalone punishment level s_{ik} gives a measure of the agent's taste for punishment absent any strategic consideration from interactions with other punishers. For ease of exposition, the following discussion focuses on the decisions of agents 1 and 2 to punish agent 3 (the target). Assume that agent 1 has a quasi-linear utility function, $u_1 = \pi_1 + v_1 (p_{13}, p_{23})$, which is strictly increasing in personal monetary earnings π_1 and weakly increasing and concave in the punishment inflicted on the target by either herself, p_{13}, or agent 2, p_{23}. Assume that agent 2 has a similar utility function.

In the one-to-one treatment, agent 1 is the only one who can punish agent 3 and agent 2 cannot do it. In other words, by design agent 1 decides on her punishment p_{13} knowing that nobody except her has the opportunity to punish agent 3. Agent *i* pays $c \cdot p_{ik}$ tokens to reduce agent *k* earnings by p_{ik} tokens, where c=1/4. She will have to balance her personal cost to punish, $c \cdot p_{13}$, versus the benefit of having agent 3 punished ($v_1 (p_{13}, 0) - v_1 (0, 0)$). In the one-to-one treatment agent 1's optimal choice is her *standalone punishment*, s_{13}=s.

Contrary to the one-to-one treatment, the most common experimental design in the literature allows for simultaneous punishments from both agent 1 and agent 2, which are then cumulated to reduce agent 3's earnings (e.g. Nikiforakis and Normann, 2008; Anderson and Putterman, 2006; Falk et al., 2005). Because of strategic considerations, agent 1's optimal choice with this simultaneous design could now be to punish anywhere between zero and her standalone punishment level, depending on how much agent 2 punishes. In the "instrumental" view, punishment by agent 1 and agent 2 are substitutes as agent 1 cares only about the fact that, as a consequence of everyone's actions, agent 3 receives a certain level of punishment.[2] For an instrumental punisher, the source of utility is the accumulated earnings reduction achieved. Hence, her marginal utility from punishment is identical whether she or agent 2 is doing the punishment (v_1 (s, 0) = v_1 (0, s)). For instance, if agent 2 has already punished *s* points, agent 1's best response is to punish zero. This assumption formalizes the concept of punishment as a second-order public good. Agents care about the free-rider getting punished but dislike having to pay the cost. When agent 1 and 2 have an identical taste for punishment (s_{13}=s_{23}=s) and choices are *simultaneous*, preferences for punishment are hard to infer from the data because any combination of punishments that sum up to *s* is an equilibrium (Varian, 1994). A coordination problem arises given this multiplicity of equilibria, which makes the interpretation of empirical evidence ambiguous. A given set of actions in the experiment may be ex-post rational or may be the result of mis-coordination among subjects.

In order to study empirically the models of instrumental punishment with the convenience of a unique equilibrium, we introduce a sequential treatment. Suppose that first, agent 1 makes her punishment choice about agent 3 and then, after learning agent 1's choice, agent 2 decides how many additional points to give. If both agents have an identical taste for punishment and that is

[2] Given the assumptions on utility, they are perfect substitutes $\partial v_1/\partial p_{13}=\partial v_1/\partial p_{23}$.

common knowledge, there exists a unique equilibrium where agent 1 punishes zero and agent 2 punishes *s* points (p_{13}=0, p_{23}=s).[3] Being the first to move puts agent 1 in a position to free-ride on the cost of punishment while enjoying a punishment equal to her standalone punishment level. In fact, any reduction in punishment by agent 1 will be exactly offset by an equivalent increase in punishment by agent 2.

If the motivations for punishment are instrumental according to the utility function above, we predict substantially different patterns of punishment in the one-to-one treatment than the sequential treatment. First, a subject will punish less, on average, in the sequential than in the one-to-one treatment. Second, if there are several potential punishers choosing in a known sequence, the first mover will rarely punish while the last mover will punish more often; hence a punisher behavior will crucially depend from the order in the sequence. Third, in the sequential treatment there are opportunities to substitute one's punishment with another's, i.e. if a subject in the sequence has already punished or is expected to punish, the others will not punish or, alternatively, reduce the amount. By design the one-to-one treatment is free from these types of strategic reasoning and subjects should ignore others' punishment choices.

3 The Experimental Design

Our design consisted of a public good experiment with three treatments within each session. There were *N*=15 participants in each session. A session lasted 24 periods and, after every period, the participants were randomly re-matched into five groups of *n*=3 individuals. To

[3] A quote from Mill (1863, ch.5): "Social utility alone can decide (among) many and irreconcilable standards of justice… I dispute the pretension of any theory which sets an imaginary standard of justice not grounded on utility… The …same motives which command the observance of morality, enjoin the punishment of those who violate them. Most of the maxims of justice are simply *instrumental* to carrying into effect the principles of justice."

familiarize with the incentive structure, in the first four periods subjects participated in a public good game without punishment opportunity. In the remaining twenty periods, subjects participated in a public good game and then observed each other's team member earnings and contribution to the public good. Following this stage, they had an opportunity to punish every other team member. For ten periods the structure was "one-to-one" punishment and for ten periods it was "sequential" punishment as it will be explained in a moment. The within subject design helps in testing the predictions, as the one-to-one treatment provides a benchmark for individual taste for punishment. At the beginning of a session, subjects were informed that the experiment had three parts and the instructions for part one were read. When part one of the experiment was completed, instructions for part two were read, and so on.[4]

In the voluntary contribution to the public good every period, each of the n subjects in a group received an endowment of y=20 tokens and made a simultaneous decision to either keep these tokens for oneself or contribute g_i tokens ($0 \le g_i \le y$) to the public good. The period monetary payoff for each subject i was given by

$$\pi_i^1 = y - g_i + a \sum_{j=1}^{n} g_j \qquad (2)$$

where a was the marginal per capita return from a contribution to the public good, a=0.6.

In the one-to-one punishment treatment, each subject simultaneously submitted two punishment requests. Only one request was actually implemented but asking for two supplied us with more data. A subject i could decrease the earnings of one other individual k in her group by p_{ik} at a private cost c=1/4. Let us designate group members with 1, 2, and 3. While group member 1 submitted one punishment request for member 2 and one for 3, one request was going

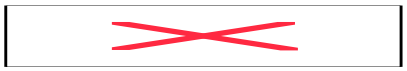
[4] Half of the sessions had a reversed order. Each punishment condition was preceded by one trial period. Instructions are available upon request. Sessions were run in May 2005.

to be carried out and the other was ignored. There were no costs for the request that was ignored. The computer simulated a coin toss and when the outcome was "heads", individual 1's target was 2, 2's target was 3, and 3's target was 1. When "tails", individual 1's target was 3, 2's target was 1, and 3's target was 2. Hence, each individual had the opportunity to punish exactly one other group member and every group member could be punished by just one other individual. These procedures were carefully explained to the subjects. At the end of the period, only the punishment requests that were carried out were made public to the group.

In the sequential treatment, each subject makes two punishment choices sequentially. There were two steps in the sequence, which will be explained through an example. In step one, subject 1 decided on the punishment of 2. In addition, 1 made a forecast about how many *additional* points of punishment subject 2 would receive in step two from 3. This forecast carried no payoff consequences and elicited the first mover's belief about the second mover's choice, which is crucial to assess the punishment motivations of the first mover. In step two, 1 decided on the punishment of 3. Before her step two decision, 1 observed how many points were assigned in step one to 3. The order of the sequence was random. In each period, every subject makes two punishment choices and both were carried out. Notice that punishment points received from the two group members cumulated and that punishment could be added but never subtracted. At the end of each period subjects observed the aggregate punishments imposed on them by the other group members, and the *aggregate* punishment carried out on *other* group members.

In all punishment conditions, subject i could request to punish any group member k by choosing, $p_{ik} \in \{0,1,\ldots,10\}$ for all $k \neq i$. The punishment received by k was subtracted from her first–stage payoff, π^1_k. Multiple punishment requests received by the same agent were

cumulated. For each point of punishment requested, subject *i* paid c=1/4 tokens, i.e., cp_{ik}.[5] In the sequential treatment, the monetary payoff for subject *i* from both stages, , can be written as:



(3)

In the one-to-one treatment, only one punishment request was selected to be actually carried out. For received punishment, subject *i*'s payoff reduction was $p_{i\,k(i)}$, where *k(i)* was the punishment request of the group member randomly assigned to subject *i*. For punishment given to others, subject *i*'s payoff was reduced by $cp_{k(i)\,i}$. The total payoff for a session was the sum of the period–payoffs for all twenty-four periods.

The experiment was conducted through computers using the "z-Tree" program (Fischbacher, 2007) with subjects anonymously interacting with each other. No subject was informed of the identity of the other group members and no communication among subjects was allowed. The payoff function, parameter values of *y*, *n*, *N*, *a* and the punishment rules were common knowledge. The public good decision was always framed in terms of contribution into a project. The punishment decision was framed as the assignment of deduction points to the other people who contributed to the same public good. We used this frame to avoid value laden terms such as "punishment" or "sanction".

The analysis is done for one-shot games. With repeated interaction and random matching (stranger), the experimental data may exhibit contagion from one period to the next. These effects may be due to (a) learning about the rules of the game, (b) learning about subject pool preferences, and (c) strategic play in earlier periods of the session. We wanted repetition to allow for (a) to take place, although it brings also (b) and (c), which are less desirable. With a perfect

[5] There was no budget constraint. For instance, a subject could purchase 10 punishment points in stage two even when stage one earnings are below 10 tokens. Negative earnings in one period were subtracted from cumulative earnings. There were no instances of negative cumulative earnings at any period in the experiment.

stranger matching, one could have avoided (c). The drawback would have been a maximum length of the session of 5 periods, which may be too short for (a). Given our logistical constraints, sessions much larger than 15 participants were impractical. There are reasons to believe that behavior of type (c) was contained. First, the probability to be in the same group in the following period was small, about 0.5%. Second, the period feedback aimed at preventing the possibility of individual reputation formation across periods; the experimental instructions explained how each subject's own contribution was always listed in the first column of his or her computer screen and the remaining two subjects' contributions were listed without subject ID in the other two columns.[6] Third, previous studies did not highlight large differences in outcome between stranger and perfect stranger matching, although not under our same design (Fehr and Gachter, 2000, 2002). If, on the contrary, most subjects believed in a considerable impact of their current action on their earnings in future periods (effect c), that would require further analysis, although our conjecture is that the qualitative predictions of the instrumental punishment view would still hold.

A total of 90 subjects were recruited among the undergraduate student population of the University of Siena via ads posted around campus. No subject had participated in public good experiments before. A total of six sessions were conducted; in three sessions the sequence was sequential and then one-to-one and in three sessions was reversed.[7] Including the reading of instructions, each session lasted about 2 hours. Payment was done privately in cash at the end of each session and was $13.90 (11 euros) per subject on average.

---

[6] At the end of each period, subjects observed individual earnings in their team with the exception of the cost of punishment given, which would have revealed information about the identity of the punishers.

[7] We conducted four Mann-Whitney tests and found no significant order effects in the data on average contributions and punishment in the sequential and one-to-one treatment (p-values ranged from 0.27 to 0.83). In the result sections the data are pooled.

4 Predictions

The canonical predictions are well known: group payoffs are maximized if each group member fully cooperates ($g_i = y$), but full free–riding ($g_i = 0$) is a dominant strategy in the contribution game. This follows from $= -1 + a < 0$. In equilibrium, subjects will contribute nothing to the public good and will not punish others, either in the sequential or in the one-to-one treatments. In fact, choosing $p_{ik}>0$ is a monetary cost that does not generate any monetary benefit in a one-shot interaction.

We consider now predictions for the punishment stage, when the agents have completed the contribution stage, know the results and face the decision about how many punishment points to give in the second stage. Consider the assumptions that interaction is one-shot, agent *i*'s utility is

$$u_i = \pi_i + v_i (g_1,..., g_n; p_{ik} + \gamma_i p_{-ik}) \quad , \quad \text{with } \gamma_i \geq 0 \qquad (4)$$

which is assumed quasi-linear in her monetary payoffs π_i, increasing and concave in the punishment given to agent *k*, $v_i'(p_{ik}) \geq 0$ and $v_i''(p_{ik}) \leq 0$. Agents' utilities may be heterogeneous although they are common knowledge.[8] The parameter γ_i measures the degree of substitutability of agent i's punishment with another agent's punishment of the same target. All predictions in this section assume that subjects are instrumental punishers, which corresponds to $\gamma_i=1$.

Prediction 1. *In the one-to-one treatment the punishment choice (p_{jk}) will be to the standalone punishment (s_{ik}) because, by design, no one else can punish the same agent, $p_{-jk} =0$.*

In the sequential treatment the more punishment others will give, p_{-ik} the lower is the optimal amount of punishment to give to agent *k*. In that case, the standalone punishment level s_{ik} is an upper bound of the choice. An instrumental punisher *i* has $\gamma_i=1$; she essentially cares about the

[8] A more complete model is $u_i = \pi_i + v_i (p_{13}, p_{23}, p_{12}, p_{32}, p_{21}, p_{31})$. In the experimental design, revenge was not possible because the information about the punishment received by the subject was revealed only at the end of a period. For instance, agent 1 could not condition her punishment strategy on p_{21} or p_{31}. The adopted specification still rules out some more complex strategies. For instance, the possibility that agent 1's punishment strategy is conditional on p_{32}.

total impact on agent *k*, 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 Bergstrom et al. (1986) and Varian (1994) developed for voluntary public good contributions.

Prediction 2 (Instrumental) *In the sequential treatment, there exists a unique equilibrium where, given a utility (4) with* $\gamma_i=1$ *and a contribution profile* (g_i , g_{-i})*, we have that,*

1) *Only one agent carries out the punishment on a target agent k.*
2) *The expected aggregate level of punishment in step one is less than in step two.*
3) *The overall punishment received by agent k is less than or equal to the maximum of all agents' standalone punishment levels,* $\max_i \{s_{ik}\}$.

The intuition for Prediction 2 is in Section 2 and a formal proof can be found in Varian (1994), where he framed it as voluntary public good contribution. Here we provide an illustrative example in a group of three members where agent 1 moves in step one and agent 2 in step two. There can be an instance where just one agent wants to punish, i.e. has a positive standalone punishment level, and other instance where both agents want to punish. In the former instance, Prediction 2 is trivial: (1) only one agent agent punishes, (2) given that the probability of moving in step one is one half, in expectation step 1 and step 2 punishment are equal, and (3) aggregate punishment is equal to the standalone punishment level of the agent wanting to punish. The case where both agents want to punish agent 3, $s_{ik}> 0$ for k=1, 2 involves more strategic reasoning. Should agent 1 punish agent 3? The best reply of agent 1, $B_1(p_{23})$, can be derived from (4) with $\gamma_i=1$:

$$u_1 = \pi_1 + v_1 (p_{13} + B_2(p_{13})) \qquad (5)$$

where the best response of agent 2 is $B_2(p_{13}) = \max\{(s_{23}-p_{13}), 0\}$. As we will see in a moment, unless her standalone punishment level is *much* higher than agent 2's, agent 1 should not punish at all because she expects agent 2 to punish.

Let us consider the three possible cases. First, when preferences are identical, $s_{13}=s_{23}$, the optimal strategy for agent 1 is to choose zero punishment because the best response of agent 2 is to choose s_{23} $B_2(0)= s_{23}$ and hence $B_1(s_{23})=0$. In equilibrium, agent 2 will bear all the cost of punishing. Notice that the order of moves solves the coordination problem that instead exists when choices are simultaneous.

Second, for the same reason as before, when agent 2 has the highest standalone punishment level, $s_{23}>s_{13}$, the optimal strategy for agent 1 is again to choose zero punishment and let agent 2 punish. Third, when agent 2 has the lowest standalone punishment level, agent 1's optimal strategy is to punish for the whole amount s_{13} only if she likes to punish *much* more than agent 2 and to punish zero otherwise. The intuition behind this strategy is that agent 1 chooses between not punishing, hence getting the preferred punishment level of agent 2, versus fully paying for his preferred level of punishment, which is higher. She will punish if the additional utility of the higher punishment is worth the cost. That happens when $u_1(, 0, s_{23}) < u_1(- cs_{13}, s_{13}, 0)$, which reduces to $\Delta_1 > s_{13}$where $\Delta_1= v_1(s_{13}) - v_1(s_{23})$. On the other hand, when preferences are similar, $\Delta_1 < s_{13}$, the optimal strategy is zero punishment as in the case of identical preferences. The sequential treatment helps in coordination as it gives clear incentives to each subject to target for punishment exclusively one other subject, who generally is the one matched in step 2.

To sum up, in the instance where both agents want to punish agent 3, then (1) one agent punishes while one does not, (2) all punishment happens in step 2, except when there is a large difference in taste for punishment between agents and the one most wanting to punish is agent 1,

which happens with probability one half, and (3) aggregate punishment is equal to the maximum standalone punishment level in the first case and sometimes in the third, and strictly less otherwise.

Consider an example with the following utility function:

$$u_i = \pi_i + \alpha_i \ln(p_{ik} + p_{-ik}), \qquad \text{with } \alpha_{ik} > 0 \qquad (6)$$

The standalone punishment level is $s_{ik}=\alpha_{ik}$. The best reply function is $B_i(p_{-ik})=\max\{0, \alpha_{ik} - p_{-ik}\}$. The indirect utility function of agent 1 is $u_1 = \pi_1 + \alpha_{13} \ln(p_{13} + \max\{0, \alpha_{23} - p_{13}\})$. In general, agent 1 will punishes if and only if $\ln(\alpha_{13}/\alpha_{23}) > 1$. For instance, when $\alpha_{13}=4$ and $\alpha_{23}=2$, agent 1's best response is not to punish; when $\alpha_{13}=6$ and $\alpha_{23}=2$ the best response is to punish.

5 Results

There are three main results.

Result 1. *In the sequential treatment, the patterns of punishment are not explained by instrumental behavior. In particular, the data do not support the prediction of a relatively higher punishment in step two than in step one. Average punishment in step two was about 10% lower than in step two.*

Table 1 shows that a contribution action in step one received on average 1.36 points of punishment compared with 1.24 points in step two (significantly lower, p-value 0.02, N=6, one-tail Wilcoxon signed-rank test). If subjects were instrumental punishers, step one punishment would be considerably lower than step two punishment.

How much lower? To address this question, we built a quantitative prediction of instrumental punishment through a simulation on one-to-one treatment data. Using these data seems

reasonable given that the underlying contribution patterns are similar between the two treatments (Figure 1, two-tail Wilcoxon signed-rank test, p-value 0.17, N=6).[9] The evolution of punishment over time has a constant trend in both treatments with no tendency to drop in the last period (two-tail Wilcoxon signed-rank test, N=6 periods 1 vs. 10, p-value 0.14 in sequential, 0.07 in one-to-one where it actually increases). The simulation relies on somewhat arbitrary assumptions and it is introduced to provide a benchmark. The simulation on one-to-one treatment data aims at understanding, *had the punishment rule been sequential*, how subjects would have punished period by period. The simulation relies on two assumptions; first, subject 1's utility is $u_1 = \pi_1 + \alpha_1 \ln (p_{13} + E[p_{23}])$; second, the expectation about step two punishment $E[p_{23}]$ is estimated with a regression on information concerning the actual contribution of the target subject in relation to others in her group, period dummies, and session dummies.[10] When the simulation results are aggregated, for every point of punishment in step two, there are just 0.35 points of punishment in step one (Table 1).[11] On the contrary, in the sequential treatment data there was no reduction of punishment in step 1: for every point of punishment in step 2, there were 1.10 points of punishment in step 1.

Result 2. *Contrary to the prediction of the instrumental model, subjects did punish in step one also when they expected the other in step two to add to their punishment. Excluding trivial cases where no punishment was given in step one nor expected in step two, about half of the decisions*

[9] Using a two-sample Kolmogorov-Smirnov test, one cannot reject that the distribution of group contributions by period in sequential and one-to-one treatments are similar (0.05 level, N=60, fifteen equally-spaced intervals of period group contributions: below 2, 4, 6, …, 30).

[10] OLS individual random effects regression on one-to-one treatment only; regressors included the average contribution of the other two persons in own group, deviation of own contribution from group average (one variable for positive and one for negative deviation) five session dummies, nine period dummies, dummy for contributions above 15 tokens.

[11] Similar punishment across steps is found also by Casari and Luini (2009) for sequential punishment within a group of five agents. A drawback of Casari and Luini (2009) is that with n=5, an instrumental punisher needs 3 steps of reasoning to compute the equilibrium. In the present study (n=3), only one step is required.

involved positive step one punishment coupled with expectations of additional step two punishment by someone else.

Result 2 is based on analysis at the level of single choices, which provide the most direct evidence on the extent (or lack) of instrumental behavior in punishment. Table 2 classifies each step one punishment choice into five cases depending on how much additional punishment is expected on the same target in step two. If no punishment is given nor expected, the situation is trivial and classified as case one. Of the remaining cases, three are compatible with instrumental punishment (2, 3, and 4) and one directly contradicts it (5). If a subject punishes in step one while expecting another subject to top it in step two (5), she could save on costs by letting the step two punisher do it all and choosing zero. Case 5 is evidence of non-instrumental behavior and it amounts to half of the non-trivial cases.[12]

This direct contradiction of instrumental punishment relies on the credibility of the elicited expectation about step-two punishment. Subjects received no additional compensation for accurate estimates. Still, the distribution of step two estimates is remarkably similar to the received step two punishment (Figure 2), and estimates have a robust, positive correlation with received step two punishment (Table 3).

Result 3. *Contrary to the predictions of instrumental punishment, the data from the sequential treatment do not show some systematic changes in patterns of individual punishment from the one-to-one treatment. In particular light and medium punishers do not scale back punishment in step one of the sequential treatment.*

Punishment choices in the one-to-one treatment reveal individual taste for punishment and were used to rank subjects based on their overall requests for punishment. Figure 3 illustrates the

[12] It is a lower bound to the amount of violations of instrumental punishment.

individual taste for punishment by the ranking of the subject within each session (thick line). Punishers 1-5 (*light punishers*) are on average responsible for 11.1% of the punishment in their session, while punishers 11-15 (*heavy punishers*) are responsible for 54.3% of the punishment. Using the above ranking, we computed individual shares of step one punishment in the sequential treatment. The instrumental prediction from the simulation is that step one punishment shares should go down for light punishers and go up for heavy punishers. That is clearly shown by the simulation illustrated by the dashed line in Figure 3. On the contrary, the actual distribution of step one punishment shows patterns that are in the opposite direction. The light punishers are responsible for 15.3% of the punishment in their session, while the heavy punishers are responsible for 46.7% of the punishment..

We draw similar conclusions when data are disaggregated by subject. Figure 4 plots the average subject punishment in step one versus step two. When a subject made the same average choice between the two steps, she would be represented as a dot on the 45 degree line. Most choices are clustered around the 45 degree line. Instrumental behavior implies that, on average, light punishers should punish more in step two and heavy punishers may be punishing less in step two, which is clearly not in line with the results shown in Figure 4. As the simulation on one-to-one data shows, strategic considerations should bring a dramatic shift away from the 45 degree line (circles, Figure 4). Summing up, the experimental results at an individual level do not support Prediction 2.

6. Possible explanations of the results

As reported in section 5, the data largely refute the instrumental model predictions. Generally, subjects do not treat punishment as a second-order public good. The open question is

then how to explain the data. We put forward and briefly discuss two conjectures, "expressive punishment" and reciprocity in punishment. While in the former conjecture decision makers are less strategic than with instrumental punishment, in the latter one they engage in more sophisticated reasoning.

Agents are expressive punishers when their utility from punishment does not depend from how much others punish and is derived only from personally inflicting the punishment. i.e. in (4) $\gamma_i=0$. For an expressive punisher, the source of utility is the action of punishing itself. As a consequence, one's punishment toward a specific target cannot be substituted with someone else's punishment that target. If agent 1 is an expressive punisher, agent 2's punishment of agent 3 has no impact on agent 1's utility ($\partial v_1/\partial p_{13}\geq 0$, $\partial v_1/\partial p_{23}=0$). The expressive model predicts the same amount of punishment in the simultaneous, sequential , and the one-to-one treatments. The concept has similarities with the "warm glow" motivation for giving in public goods (Andreoni, 1990). While Prediction 1 holds also for expressive punishers, new predictions are in order for the sequential treatment:[13]

Prediction 3 (Expressive). *In the sequential treatment, there exists (trivially) a unique equilibrium where, given a utility (4) with* $\gamma_i=0$ and *a contribution profile* (g_i , g_{-i})*, we have that,*

1) *Both agents may punish and will do it with the same frequency as in the one-to-one treatment.*
2) *The expected aggregate levels of punishment in step one and step two are identical.*
3) *The overall punishment received by agent k is the sum of the other agents' standalone punishments,* $\sum_{i\neq k} s_{ik}$.

[13] Alternative labels are "outcome-based utility" for the instrumental model and "action-based utility" for the expressive model.

Prediction 3 relies on subjects not changing punishment choices according to which step they are in. For instance, given that a subject is equally likely to be assigned to step 1 or to step 2 , in expectation there will be equal punishment in both steps even with heterogeneous tastes for punishment. Prediction 3 fits Results 1, 2, and 3 better than Prediction 2. It also makes additional predictions about the comparison of one-to-one and sequential treatment results. For instance, the fraction of punished actions is comparable between treatments (76.1% in one-to-one and 72.8% in sequential, two-tail Wilcoxon signed-rank test, N=6, p-value 0.17). On point 1 of Prediction 3, we report a lower frequency of multiple requests to punish in the sequential (38.8%) than in the one-to-one treatment (46.4%). While the sequential result is closer to 46.4% than to the instrumental punishment prediction of nearly 0%, it indicates some degree of substitutability in the punishment requests across subjects. In other words, in (4) the data suggest a $\gamma_i>0$.

On point 3 of Prediction 3, we report that the sum of requests in the one-to-one treatment averages 3.32 points per each contribution choice, which should be the same under the assumptions of equivalence in contribution patterns between treatments (Table 1). Instead the sum of punishment for the sequential treatment was lower (2.61, i.e. 78.6% of the one-to-one data), which again suggests that there is some degree of substitutability in giving punishment. Hence, Prediction 3 point 3 is not fully supported. On the other hand, the instrumental punishment simulation yielded an average punishment of 2.12 (66.5%). Overall, the data show considerable, although not full, support for expressive punishment.[14]

We now discuss the reciprocity in punishment conjecture as an alternative explanation of the results. A subject may punish in step 1 in order to show leadership and have the other subject

[14] For a heavy punisher, the punishment expense could be higher in the sequential than in the one-to-one treatment. If the demand for punishment depends from the income of the punisher, it may lead to a lower demand for punishment in the sequential treatment. If an income effect exists, though, one can presume not an effect from period income but from the expected session income. On the contrary, Figure 5 shows that the share of punishment requested by heavy punishers in the sequential treatment is very similar to the one-to-one treatment.

punishing as well in step 2. Should she not punish, the step 2 mover observes it and may retaliate by not punishing either. This conjecture would provide a reason of why both subjects punish despite the sequential design. We test whether a reciprocal response involved a "less-than-usual" step two punishment when step one punishment was not "adequate" and a "more-than-usual" step two punishment otherwise. We take data from the one-to-one treatment as benchmark for usual punishment and no punishment as proxy for inadequate punishment. In Figure 5 one can see from the exercise done with one-to-one data that when a person did not punish (assume it to be step 1), in 53.1% of the cases also the other person did not punish (assume it to be step 2).

Instead, when one person gave two points of punishment then only in 33.9% of the cases the other person did not punish. This decline of no punishment was not due to reciprocity but simply reveals a similarity of punishment norms among subjects. We take it as benchmark of "usual" punishment and measure against it the data from the sequential treatment. The solid line with round dots in Figure 5 reports the fraction of step two choices without punishment for the cases when in step one a subject punished 0, 1, 2, or 3-10 points. The line for the sequential treatment is not "steeper" than the one-to-one treatment but roughly parallel. This evidence is squarely against the reciprocity in punishment conjecture. This analysis does not reveal any leadership effect induced by step 1 punishers because step 2 punishers do not react differently to observed zero vs. positive punishment. Although not exhaustive, this follow-up analysis reveals more support for expressive punishment than for reciprocity in punishment.

7. Conclusions

A key question about human societies is how social norms of cooperation are enforced. Subjects who do not obey a social norm are often targeted by their peers for punishment. While this

provides an answer about enforcement, many aspects of what motivates punishers are still unclear. These aspects are quite relevant to assess the social desirability of informal peer punishment versus legal punishment as alternative ways to enforce norms. They are also relevant for the general theoretical debate about what are the strongest other-regarding motivations in individual economic choices.

We examine these questions through an experiment with a design that provides an original and insightful viewpoint for uncovering subjects' motivations. There is a generic agreement in the literature that emotions play an important role in motivating punishers, but very few details on how they may be formalized in a model (Vyrastekova et al., 2008; Falk et al., 2005; Xia and Houser, 2005). A novel aspect of this paper is to formalize the roles of emotions and strategic behavior in peer punishment and test them empirically. Initially, we put forward a model of *instrumental* punishment, which takes the widespread view that agents care only about the income consequence of their actions. In the experiment subjects first contributed to a public good and then had an opportunity for peer punishment. In one treatment punishment choices were sequential and in the other, one-to-one treatment, there was no strategic element in punishment. The instrumental model makes predictions about the timing, magnitude, and target of punishment. No predictions of the instrumental model find solid support in the data.

On the contrary, we find that a model of *expressive* punishment, where utility comes from the personal act of punishing, provides a better explanation for the results than a model of instrumental punishment. The expressive model may reflect a strong role of emotion in driving punishment as well as a role for preserving the identity of the agent as norm follower (Charness et al., 2007), which somewhat disregards the incentive consequences of the accumulation of punishment from everyone in the team. We did not test whether emotions shaped directly the

utility function or simply interfered with the ability of subjects to reason strategically after the norm violation.

The main finding is that subjects do not treat peer punishment as a second-order public good and hence the alignment of individual motivations to punish and social welfare would be purely accidental. Subjects' most important goal is not to provide incentives for the free-rider to contribute. One implication of the expressive model is that in large groups there could be an excess of peer punishment. Large groups may be better off by providing alternative, less destructive channels to express emotions (Xiao and Houser, 2005), by letting people select institutions (Guererk et al., 2006; Sutter et al., 2008), or by appropriately restraining peer punishment (Casari and Plott, 2003).

Although peer punishment is the only option in many situations, societies may otherwise benefit from adopting legal punishment systems. One advantage over peer punishment is that legal systems explicitly aim at deterring free-riding through an overall sanction proportional to the crime. Another advantage is that they follow strict formal procedures in an attempt to isolate punishment decisions from emotional responses to the crime. The evidence from this study suggests that, in important ways, peer punishment is not guided by the aim of deterring crime but instead by the personal satisfaction of taking revenge.

References

Anderson, C. M. and Putterman, L. (2006) Do Non-strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism. *Games and Economic Behavior* 54, 1, 1-24.

Andreoni, J. (1990) Impure altruism and donations to public goods: a theory of warm-glow giving, *Economic Journal* 100, 464-77.

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.

Carpenter, J. and Matthews, P.H. (2009) What norms trigger punishment, *Experimental Economics* 12, 272–288.

Casari, M. and Luini, L. (2009) Group cooperation under alternative peer punishment technologies: an experiment, *Journal of Economic Behavior and Organization* 71, 273–282.

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.

Falk, A., Fehr, E., Fischbacher, U. (2005) Driving Forces behind Informal Sanctions, *Econometrica* 73, 6, 2017-30.

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

Fischbacher, U. (2007) Z-Tree: Zurich Toolbox for Readymade Economic Experiments. Instructions for Experimenters, Experimental Economics, 10, 2, 171-178.

Gächter, S., Hermann, B., Thöni, C. (2005) Cross-cultural differences in norm enforcement, *Behavioral and Brain Sciences* 28, 822-23.

Guererk, Oezguer, Bernd Irlenbusch, and Bettina Rockenbach. The competitive advantage of sanctioning institutions. *Science* 312, 5770, (2006): 108-111.

Mill, J.S. (1863) *Utlitarianism*, Longmans.

Nikiforakis, N. and Normann, H.-T. (2008) A Comparative Statics Analysis of Punishment in Public Goods Experiments, Experimental Economics, 11, 4, 358-369.

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.

Sutter, M., Haigner, S., and Kocher, M. (2008) Choosing the carrot or the stick? – Endogenous institutional choice in social dilemma situations, Working Papers 2008-07, Faculty of Economics and Statistics, University of Innsbruck.

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

Vyrastekova, J., Funaki, Y., Takeuchi, A. (2008) Strategic vs Non-Strategic Motivations of Sanctioning, Tilburg University, Center for Economic Research, working paper.

Xiao E, Houser D (2005) Emotion expression in human punishment behavior, *Proceedings of the National Academy of Sciences for the United States of America* 102, 20, 7398-7401, May 17.

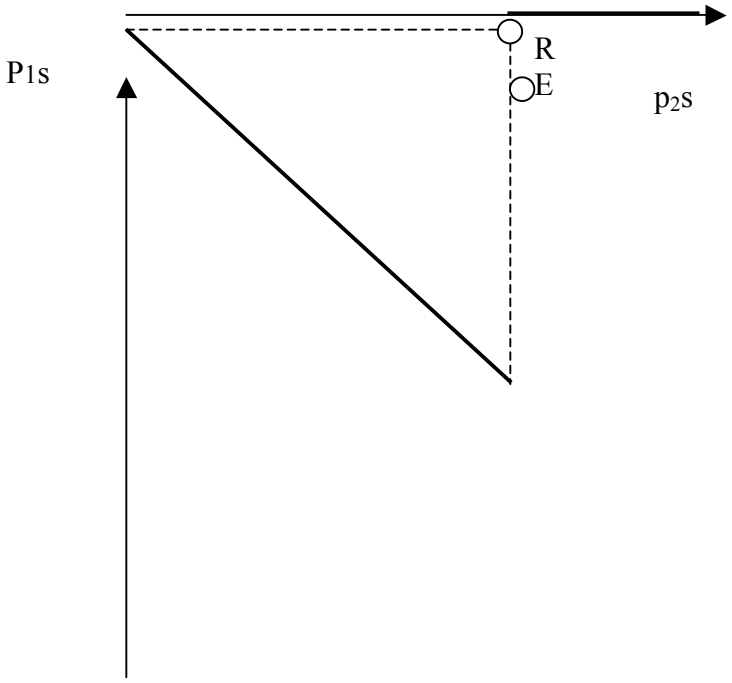


Figure 1: Unique punishment equilibrium under sequential moves

Agent 1 punishes 0 points and agent 2 punishes *s* points (p_{12}, p_{23}=s, point R)

(hp: Both agents have an identical taste for punishment that is common knowledge)

Figure 2: Contribution in the punishment treatments

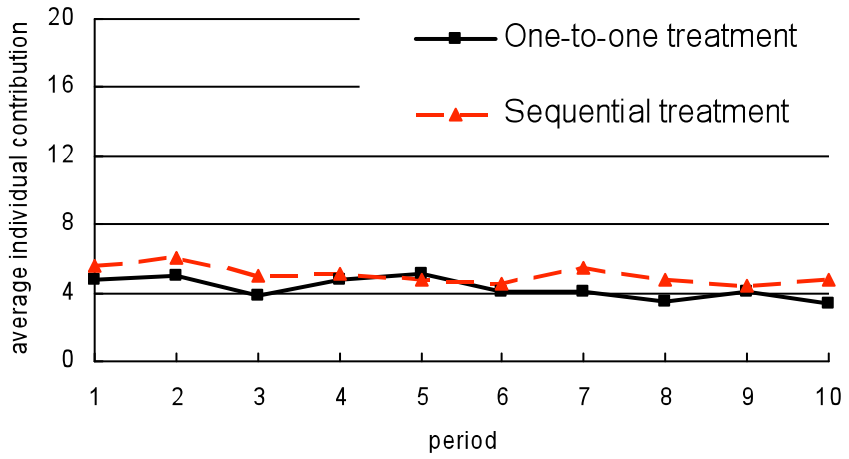
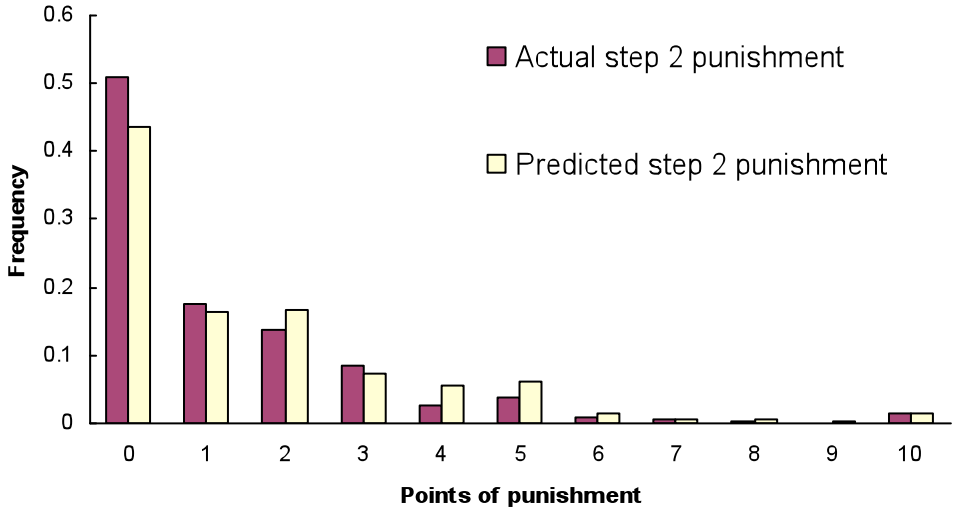
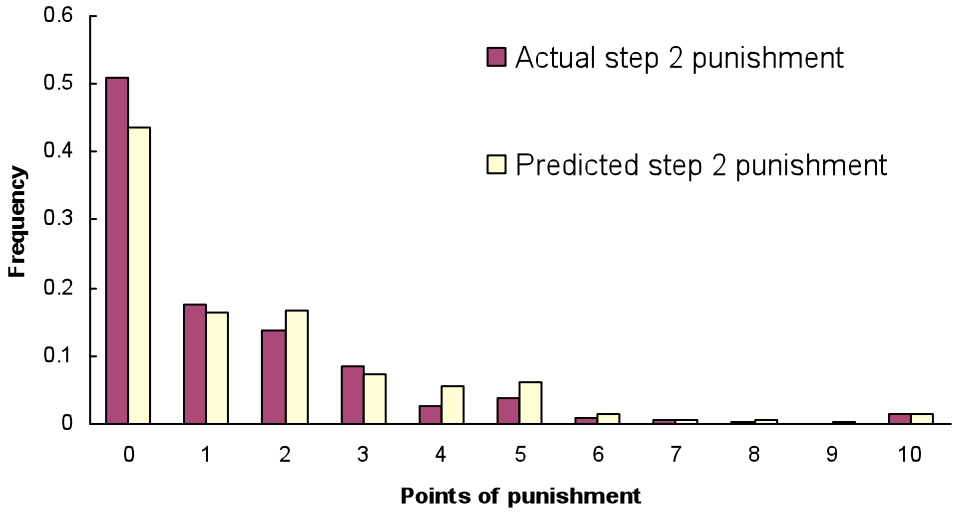


Figure 3: Distribution of actual and predicted punishment points by level



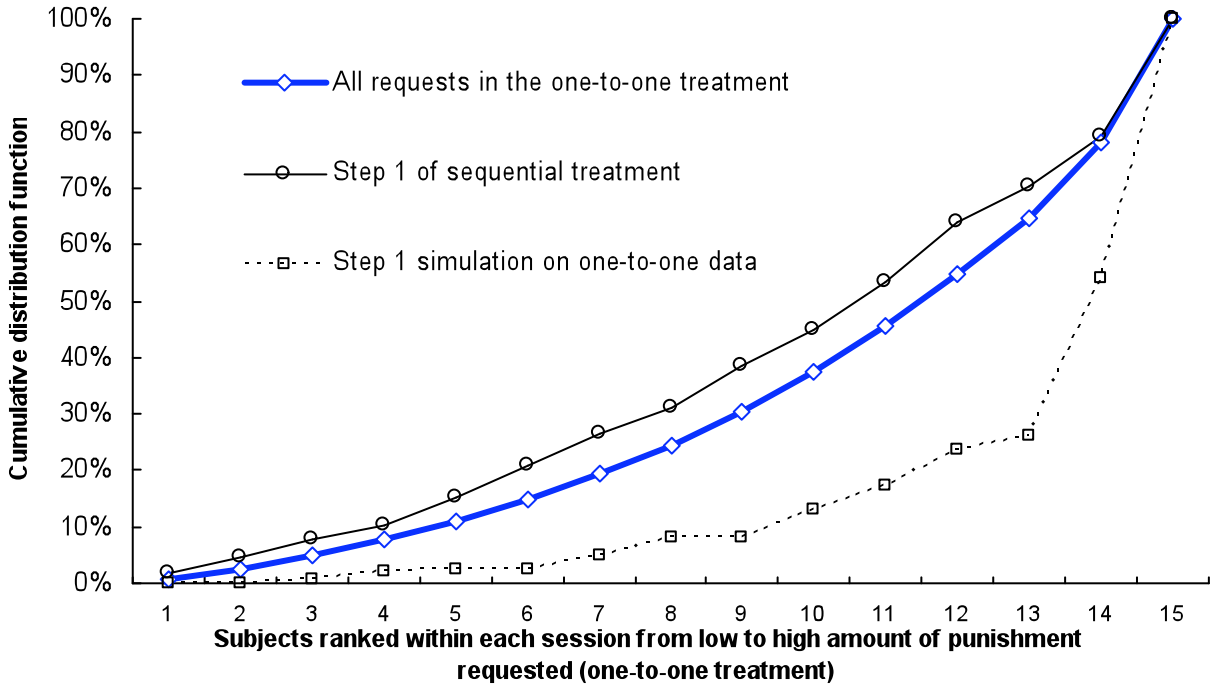
Notes: N=900, Overall distribution by punishment level

Figure 4



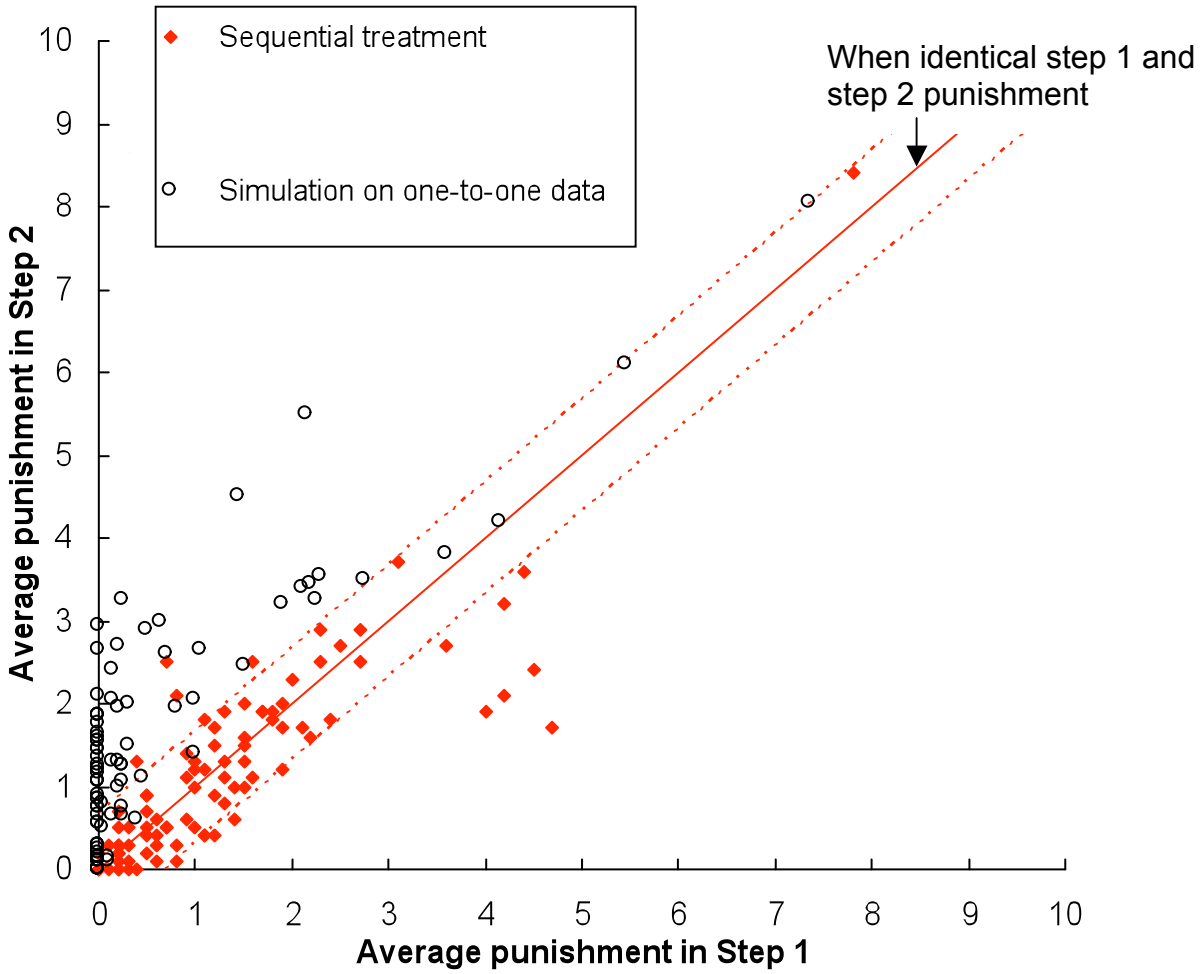
Notes: N=900, Overall distribution by punishment level

Figure 5: Subjects ranked by taste for punishment



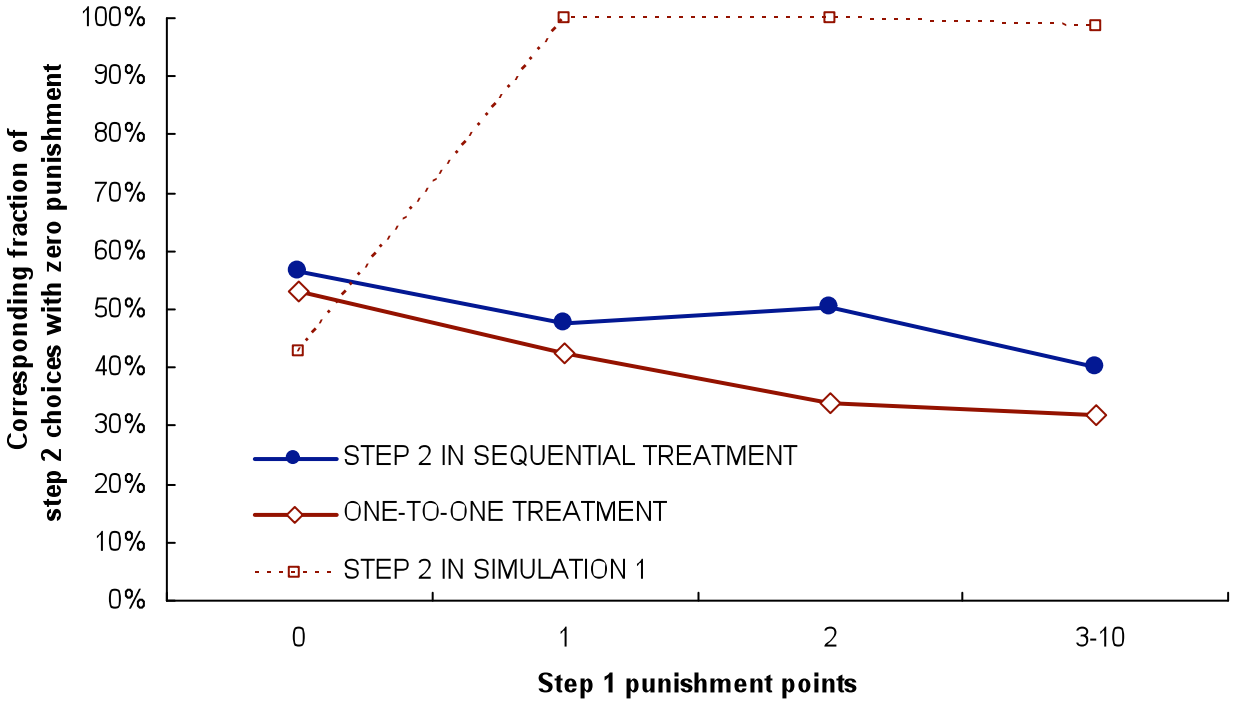
Notes: The vertical axis reports shares of total punishment in a session; the figure reports averages of all six sessions. Step 1 simulation is done using one-to-one treatment data.

Figure 6: Do subjects punish differently in step 1 versus step 2?



Note: If punishment is instrumental, the simulation data and the sequential data should exhibit similar patterns. N=90. Each point represents an individual, i.e. average amount of punishment requested. The solid line is the 45 degree line; the dotted line indicates one standard deviation of the individual (step one – step two) difference.

Figure 7: Punishment norms and reciprocity in punishment



Notes: The one-to-one treatment line is computed as the average of two scenarios. One is when the first request is considered step one punishment while the other is when the second request is considered step one punishment. The dashed line is the instrumental simulation on one-to-one data that was also used in the other figures and that is here for comparison.

Table 1: Punishment requested

Sequential treatment		
Step 1	1.37	
Step 2	1.24	
Total		2.61
Predicted Step 2 punishment	1.59	
One-to-one treatment		
Carried out punishment	1.66	
Requested and not carried out	1.66	
Total		3.32
Instrumental simulation on one-to-one data		
Step 1	0.56	
Step 2	1.56	
Total		2.12

Note: average punishment points requested, cp_{ik}

Table 2: A classification of step one punishment requests in the sequential treatment

Case	Punishment given in step 1	Prediction about additional punishment in step 2	Description	Number of obs.
1	zero	Zero	No punishment done nor expected	277 (31%)
2	+	Zero	Either the subject is the only one wanting to punish OR is a heavy punisher who jumps in step 1	116 (13%)
3	zero	+	Either the subject will not punish in any case OR will let the other do the punishment for her	156 (17%)
4	+	+	The expected sum is greater than 10; The subject needs the cooperation of the other to reach the desired level of punishment	41 (5%)
5	+	+	The expected sum is less than or equal to 10; The subject punishes expecting that the other will punish as well	310 (34%)
			Totals	900 (100%)

Notes: + stands for a strictly positive amount of punishment; cases 2, 3, 4, 5 are labeled as "non-trivial."

Table 3: Relation between predictions and punishment in step two

Dependent variable: Received step two punishment	(1) session and period dummies	(2) session dummies	(3) no dummies
Prediction about additional step two punishment	0.1076*	0.1131*	0.1597***
	(0.0588)	(0.0587)	(0.0583)
Constant	0.1596	0.1432	-0.3555**
	(0.4667)	(0.3015)	(0.1695)
Observations	900	900	900

*Notes: Tobit regression; * significant at 10%; ** significant at 5%; *** significant at 1%; Standard errors in parentheses; session and period dummies omitted from table.*