Cooperation under alternative punishment institutions: an experiment

Casari, Marco; Luini, Luigi

Postprint / Postprint
Zeitschriftenartikel / journal article

Title: Cooperation Under Alternative Punishment Institutions:An Experiment

Authors: Marco Casari, Luigi Luini

PII: S0167-2681(09)00090-0
DOI: doi:10.1016/j.jebo.2009.03.022
Reference: JEBO 2362

To appear in: *Journal of Economic Behavior & Organization*

Received date: 3-12-2006
Revised date: 30-3-2009
Accepted date: 30-3-2009

Please cite this article as: Casari, M., Luini, L., Cooperation Under Alternative Punishment Institutions:An Experiment, *Journal of Economic Behavior and Organization* (2008), doi:10.1016/j.jebo.2009.03.022

COOPERATION UNDER ALTERNATIVE PUNISHMENT INSTITUTIONS: AN EXPERIMENT *

Marco Casari
University of Bologna

and

Luigi Luini
University of Siena

Abstract

While peer punishment has been shown to increase group cooperation, there is open debate on how cooperative norms can emerge and on what motives drive individuals to punish. In a public good experiment we compared alternative punishment institutions and found (1) higher cooperation levels under a consensual punishment institution than under autonomous individual punishment; (2) similar cooperation levels under sequential and simultaneous punishment institutions.

JEL C91, C92, D23.

Keywords: public goods, peer punishment, social norms, team production, experiments.

* Corresponding author: Marco Casari, Department of Economics, Università di Bologna, Piazza Scaravilli 2, 40126 Bologna, Italy, Tel: +39 051 209 8662, Fax: +39 051 209 8493, marco.casari@unibo.it; 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

Cooperation in groups is more likely to arise when peers can punish each other. The first wave of experiments on informal sanctions proved this point using a special punishment rule (Ostrom et al., 1992, Fehr and Gächter, 2000). 1 Since then scholars have been experimenting with variants in the form of punishment that reflect the variety of unstructured interactions that could characterized a social group (Andreoni et al., 2003, Decker et al., 2003, Denant-Boemont et al., 2007, Xiao and Houser, 2005, Camera and Casari, 2009, Casari, 2005, Henrich et al., 2006, Guererk et al., 2006).

Through a novel experimental design we tackle the issues of how a cooperative norm emerges and what motives drive individuals to punish (Camerer and Fehr, 2006, Boyd et al., 2003). We show that the emergence of norms of cooperation critically depends on the form of peer punishment available. Inside a group there may be competing individual norms of behavior and the punishment institution available have a fundamental role in composing those norms and hence generating the group outcome. For a given group, either a cooperative norm or a free-riding norm could emerge depending of the specific punishment institution. In the laboratory one can supply one punishment institution at a time and precisely isolate its effect on group cooperation. When everyone can punish without constraints (Baseline treatment), group cooperation is more easily crippled by a minority of individuals that are spiteful or obey to a free-riding norm. This has been documented by a number of studies (Gächter and Herrmann, 2006, Houser et al., 2005, Cinyabuguma et al., 2004). We report about a form of peer punishment, the Consensual institution, which endogenously censors the free-riding norm and greatly enhances group performance. Under the Consensual rule a request to punish is ignored when punishment toward a specific group member is requested by one agent only. Hence, peer punishment is carried out only when there is a coalition of two or more agents that share the

same norm. We document this "consensus dividend," i.e. an overall high group performance, whenever a coalition is needed to carry out peer punishment. This virtuous institution represents an interesting variant of informal punishment as the experimenter never imposes any norm on who can or cannot be punished; these norms emerge endogenously from the interaction within the group. Groups with a social dynamic that resemble to this punishment rule can accrue a consensual dividend.

A second result is that group members seem uninterested in coordinating their punishment actions. When information is given about how much others have already punished a group member, the subject does not adjust her punishment request accordingly. We call this surprising handling of such additional information the "coordination puzzle." This result provides insights into the motivations for punishment because it suggests that the punisher derives her utility from the act of punishing in itself and not from achieving, in conjunction with other punishers, a total amount of punishment that would discourage the free-rider. If that is the case, group interaction with peer punishment will achieve aggregate efficiency only by accident.

This paper is structured into four Sections. In Section 1 we describe the experimental design and the predictions. Aggregate results are presented in Section 2, while individual punishment decision results are presented in Section 3. Conclusions follow in Section 4.

1. The Experimental Design

Our design consists of a voluntary contribution public good with the opportunity to engage in peer-to-peer punishment. The experiment includes three treatments with distinct punishment rules, Baseline, Sequential, and Consensual.[1] There are N=20 participants in each session. In every period the participants are randomly partitioned into four groups of n=5 individuals. A

[1] The instructions for the Consensual treatment can be found in Appendix.

session comprises two parts for a total of twenty periods of a public good game. While part 1 is a simple voluntary contribution to a public good game, in part 2 there is also a punishment opportunity.

Irregardless of the treatment, in part 1 (periods 1-10) there is no punishment opportunity. Every period each subject *i* receives an endowment of 20 tokens and chooses to contribute $g_i \in [0, 20]$ tokens to a group project along with other *n*-1 other subjects in her group. All contribution decisions are made simultaneously. Period earnings for subject *i* in periods 1-10 are:

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

where a=0.4 is the marginal per capita return from a contribution to the public good. At the end of each period subjects are informed about the total contribution Σg_j to the project as well as contributions and earnings of every member in their group. In the stage game, full free–riding (g_i = 0) is the dominant strategy. 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 =20) because $\partial \sum_{i=1}^{n} \pi_i^1 / \partial g_i = -1 + na > 0$.

At the start of the session, we announce that the experiment has two parts but explain the rules just for the first part.[2] No subject is ever informed about the identity of the other group members. No communication among subjects is allowed. After each period, subjects are randomly and anonymously re-matched in groups of five and the probability that an agent is re-matched with the same four people is less than two percent.[3]

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

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

In part 2 (periods 11-20) subjects have an opportunity to punish others according to rules that differ in the three treatments. Each period includes two stages. At stage one, subjects simultaneously choose contribution levels. At stage two, subjects are informed about the individual contributions of all other group members. Moreover, a subject *j* can request to punish any of her group members *i* by assigning punishment points $p^i_j \in \{0,1,\ldots,7\}$. More precisely, every subject faces four decisions of assigning punishment points, one for every other person in her group; a subject cannot punish people outside her group. Punishment rules differ by treatment.

In the Baseline treatment punishment choices are simultaneous. 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. This is a common protocol in the experimental literature, adopted, for instance, by Fehr and Gächter (2002).[4] Period earnings for subject *i* in periods 11-20 are:

$$\pi_i = \pi_i^1 - 3\sum_{k \neq i} p_k^i - \sum_{k \neq i} p_i^k \qquad (2)$$

Session earnings were the sum of earnings in all periods. When deciding on punishment, the computer screen shows a table with each subject's own contribution always listed in the first column and the remaining four subjects' contributions listed in the other four columns without subject identifiers. This feature accomplishes several goals: it prevents the formation of individual reputations across periods; it makes it difficult to delay punishment to following periods; it makes it difficult to punish for revenge. At the end of a period, subjects can observe

[4] The fine-to-free ratio is constant at 3 to 1. Period earnings of a subject can be negative, although in the experiment that event was infrequent. When ignoring the punishment given to others, the frequency was 3.3% in periods 11-20. Cumulative earnings were always positive.

the aggregate punishments imposed on them by the other group members, and the *aggregate* punishment imposed on *other* group members but do not know who requested such punishment.[5]

In the Sequential treatment the only difference from the Baseline treatment is the timing of the punishment decisions. At stage one, subjects simultaneously choose contribution levels. Instead, in stage two a subject can punish each one of the other $(n-1)$ group members in $(n-1)$ separate steps. In step one each subject places a punishment request on a person. In step two, a subject places a punishment request on another person knowing how much punishment has been given to that person in step one by someone else. The process continues for four steps until a subject has had the opportunity to target every other member in her group. To summarize, at step k agent i can punish just agent $j(k)$; the order of punishment decisions is random. Punishment points can be added but never subtracted.

When punishing, a subject knows at what step she is in the sequence and also the cumulative *aggregate* punishment imposed on each *other* group members up to the previous step (Varian, 1994). Hence, in the Sequential treatment a subject receives more detailed information about punishment than in the Baseline treatment because she can see both the end-of-period sum and some disaggregated statistics about the individual components of this sum. However, she is not informed about the amount of punishment she has personally received until the end of the period. This provision is meant to prevent, as much as possible, a subject from using punishment to payback others for their requested punishments.

In the Consensual treatment both contribution and punishment decisions are simultaneous. The peculiar aspect of consensual punishment is that an agent is punished only if at least two agents

[5] 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 Gächter, 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.

requested it. In practice, only a coalition of 40% of group members or larger is allowed to punish a person. When there is just one request to punish agent *i*, it has no effect. More precisely, agent *i* keeps her stage one earnings without reduction and will not know of the punishment request. Moreover, there is no cost for requesting punishment if it is not carried out. Period earnings for subject *i* in periods 11-20 are:

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

where $K(i) = 1$ if $\left[\sum_k I_{\{i,k\}}\right] \geq 2$ and $K(i) = 0$ otherwise. The function $I(i,k)$ equals one when agent *k* requests to punish agent *i*, $p^i_k > 0$, and equals zero otherwise. To carry out the punishment, the consensus must be who is the target and not necessarily the exact severity of the sanction.

The canonical predictions for the experimental conditions just outlined are well known. If subjects apply the backward induction logic, the 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.

2. Aggregate 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

twelve sessions. Payment was done privately in cash at the end of each session and averaged 12.40 euros per subject.[6]

In this section we present results on aggregate cooperation and net payoff (Results 1-3) while in the next session those concerning individual decisions to punish (Results 4-5).

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 rises 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 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 1 and Figure 1. Without a punishment opportunity the average individual contribution across all treatments is 3.31 tokens. When the opportunity to punish is introduced, the average individual contribution across all treatments is 5.77. A Wilcoxon signed ranks test shows that this difference in contributions is significant at the two percent level (N=12). These average values hide a declining trend when there are no opportunities to punish - from 5.92 tokens in period one to 1.82 in period ten. With punishment opportunities there is a "jump" in period eleven when there is an average contribution of 5.21 tokens and an ascending trend 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 (N=12).

[6] The average payment is $14.50 at the October 2003 conversion rate. This amount includes the show up fee that was 3 euros for the four sessions conducted before October and 5 euros afterwards. A "Token" was worth 0.02 euros. Each session lasted between 1 hour and 50 minutes and 2 hours and 30 minutes including instructions reading.

Besides these common patterns, each punishment rule shows remarkable peculiarities. Overall contributions under a Consensual rule are substantially higher than in the other two (8.46 vs. 4.46 Baseline and 4.38 Sequential).[7] Moreover, while the time trend is increasing for the Consensual rule (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). Another way to acpture these dynamics is to compare net earnings considering punishment expenditure and costs over time (Result 3). There exists a relative payoff loss within a treatment if the net earnings in period t under the punishment condition are lower than the earnings in period t under the no punishment condition.

RESULT 3: *The punishment condition initially caused a relative payoff loss. In the Baseline and Sequential treatments the relative payoff losses remained throughout all periods, although they became smaller over time. Payoff losses and gains differ by treatment especially when considering the last periods. In the final period of the Baseline and Sequential treatments the relative payoff loss was roughly 20 percent. In the final period of the Consensual treatment the relative payoff gain was 13 percent.*

When 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.

Interestingly, the high cooperation level of the Consensual treatment was achieved with the lowest level of punishment among all treatments. This is the key point that accounts for its superiority in terms of group net earnings. Let us define the "punishment rate" as the number of

[7] There is considerable variance in the effect of the consensual punishment rule. In particular the jump in contribution is driven by two sessions out of four.

punishment points assigned to a particular contribution action. The average punishment rate is $\sum_{t=1}^{10}\sum_{j}^{n}\sum_{k\neq j}^{n} p_{k,t}^{j}/10n$ for Baseline and Sequential and $\sum_{t=1}^{10}\sum_{j}^{n} K_j/\sum_{k\neq j}^{n} p_{k,t}^{j}/10n$ for Consensual. The average punishment rate was 1.70 in the Consensual compared to 2.47 in the other two treatments (Table 2). 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 29.4 percent higher than our reported rate.

What cries for an explanation is how a lower threat of punishment observed in the Consensual treatment could provide not weaker but *stronger* incentives to cooperate that in the other treatments. The reason is that the Consensual rule endogenously filtered out the anti-social norm of a minority that was targeting cooperators, thus enhancing the incentives to cooperate. While less than one out of every ten requests to target full free-riders was censored, more than seven out of ten attempts to punish strong cooperators with contributions (15,20] were blocked (Table 2). We will come back to this point in the next section. The Baseline and Sequential rule instead allowed a minority to freely harm strong cooperators and hence lower incentives for cooperation.

What stands out in the analysis of group cooperation levels across treatments is the superiority of the Consensual rule. This rule generated punishment costs 10 percent lower that the Baseline rule and realized a contribution level 90 percent higher.

3. Individual contribution and punishment decisions

A data analysis at the individual level gives additional insights into the performance of peer punishment institutions. We first present Results 4 and 5 about the frequency of multiple

punishment requests on the same target and then detailed econometric models on punishment and contribution choices.

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 3. To interpret this result, it is important to consider that in about 92% of the instances one group member could carry out single-handedly the whole punishment. A subject could distribute up to seven points of punishment to another subject. [8] A total of eight or more points were distributed to a subject only in 8.2% of the cases. It follows that the multiplicity of requests to punish the same agent is not a response to the need to punish free riders more severely because almost always one agent could have done it alone.

We consider two possible explanations for this multiplicity of punishment requests for the same target in the Baseline treatment:

(a) Under the interpretation that peer punishment is a "second order public good," Result 4 could be evidence of a coordination failure. Assume that some agents are willing to punish the free riders if no one else does. According to the "second order public good" view those agents derive utility from having an agent punished and hence are willing to pay a private cost to punish. An agent of the type above would happily free ride on punishment if she knows that somebody else will punish. For instance, if it is common knowledge that agent *i* wants to punish a given target for 3 points and agent *k* wants to punish for 6 points, then agent *i* can free ride on the punishment of agent *k*. Different

[8] Seven points of punishment reduces earnings by 21 tokens, which implies an earning reduction between 40% and 105%.

choices may occur if there is uncertainty about agents' punishment preferences, which can explain Result 4 as a coordination failure.[9]

(b) In another interpretation subjects' decisions to punish do not depend on how much others punish the same subject. Stated differently, if a subject gains utility only from her personal punishment action then she does not care about the total amount of punishment received by the targeted subject. This preference structure could describe a strong emotional drive in the motivations for punishment. In that case no strategic element would enter into the punishment decision and the multiplicity of punishment requests would cease to be a puzzle. It would simply reflect the plurality of subjects in each group with a preference for punishment.

Under the "emotional" interpretation of punishment decisions (b), the procedural differences between Baseline and Sequential would be irrelevant for punishers. As described in the next result, the data provide more support for (b) than for (a), the "second order public good" interpretation.

RESULT 5: *In the Sequential treatment there is no improvement in coordination in punishment in comparison with the Baseline treatment. In particular, we observe across treatments similar frequencies of multiple requests to punish the same subject and similar distributions of total punishment received by free riders.*

The similarity in the multiplicity of requests to punish is detailed in Table 3. According to coordination in punishment decisions is easier in the Sequential than in the Baseline treatment because later movers in the sequence have additional information on the punishment already assigned to the target. As a consequence, one may expect in the Sequential data a lower number

[9] Under (a) there is a parallel between *consensual* peer punishment and contributions to a threshold public good with refunding. While in public goods experiments the threshold is generally on the aggregate contribution level, here the threshold is in terms of number of punishers, irregardless of the level of punishment requested.

of group members targeting the same agent than in the Baseline. Empirically, that does not seem to be the case (Table 3).

We cannot rule out that the information acquired throughout the steps of punishment enters into the decisional process. Yet, the data suggest that this information is not used according to the "second order public good" interpretation (a). Table 4 presents a detailed econometric model on why subjects punish and allows a comparison of punishment expenditure when a subject was alone in punishing the target with the cases when everyone in the group punished the target. In step 1, there is no significant difference in the estimated coefficients. The subject may have not cared that other would also punish in future steps (interpretation b) or may have been unable to predict future punishment choices (interpretation a). In step 4, if nobody punished the target before, punishment is significantly higher than if three people already punished the target. This evidence is compatible with interpretation (a), because one may conclude that subject cared whether others punished the target. The evidence though is weak. First of all a significant difference between the two coefficients was found also for the other treatments. Moreover, in a modified step 4 regression (unreported) with independent variables for two, three, and four people punishing the target, one cannot reject the hypothesis that they are all equal (p-value=0.27).

The evidence on total punishment received by free riders provides further support in the same direction. When two or more subjects in a group are willing to punish there could be a problem in coordinating punishment. If there is an improvement in coordination, one would expect to see in the Sequential results less variability in the punishment received by free riders, i.e. a reduction in the number of free riders escaping punishment or receiving extremely high punishments. Also this conjecture about an improvement in coordination is not supported in the data. We present

data relative to groups where two or more members contributed positive amounts *and* where at least one complete free rode (zero contribution). These situations are very common as they account for 68.1% of the groups in the Baseline and 70.6% in the Sequential treatment. Typically, the complete free riders received a heavy punishment, an average of 4.83 points in the Baseline and of 4.32 points in the Sequential. The actual punishment did vary widely in level from 0 to 19 points. Yet, the empirical distributions of the punishment points targeting complete free riders are surprisingly similar between Baseline and Sequential treatment. A Kolmogorov-Smirnov test for equality of distribution functions cannot reject the equality hypothesis (p-value of 0.34, N=183, 219).[10] We conclude that the additional information provided in the Sequential compared to the Baseline treatment did not significantly change either the level or the dispersion of punishment decisions (the standard deviation in punishment points actually grows slightly from 3.46 in the Baseline to 3.85 in the Sequential). To sum up, there is a puzzle here for interpretation (a), as the additional information available in the Sequential treatment was not used accordingly.

The remaining of this section further discusses the motivations of punishers, the effects of sanctions on cooperation levels, and the peculiarity of the consensual treatment. Why do people punish? There are three main findings common to all treatments from the regressions in Table 4. First, the contribution level of the target matters. Punishment is heavier for the lowest contributor in the group and lighter for the highest contributor. Previous studies find a similar result while employing as regressor the target's contribution minus the group average contribution. Our specification avoids any hypothesis on the functional form of the relationship. Second, when

[10] The test assumes observations are independent. If there is dependence the result of no significant differences may be even stronger.

others are punished in the previous period, this encourages the subject to punish more. This "imitation" effect in punishment is stronger than the "blind revenge" effect of punishing after having been personally punished in the previous period. Third, the contribution level in the part without sanctions is a poor predictors of punishment choices. In particular, we do not find subjects who are cooperating type and who will punish when the opportunity is given. Subjects who free-ride without sanction, may also be willing to engage in punishment when given a chance. Requests to punish are context-specific and depend from the subject's relative contribution within the period.

What is the effect of sanctions on contribution choices? From the regressions in Table 5 there are four main findings common to all treatments. First, the immediate consequence of receiving punishment is to *lower* contribution levels, which runs contrary to our intuition. This effect is highly significant and concerns punishment received both in the previous and in the second previous periods. Second, after requesting punishing in the previous period the subject also lowers her contribution level. This behavior could be interpreted as a trigger strategy following a norm violation. The subject punishes directly through punishment points and indirectly through withholding future cooperation. Third, punishment boosts cooperation because the subject observes that others, especially free-riders, have been punished. *This seems to be the main positive effect of punishment on contribution.* The exception is when the highest contributor in the group was punished in the previous period, which has a negative impact on contribution levels. We label the latter behavior "perverse punishment" and will return to it in a moment. Forth, the average contribution in the part without sanctions is a good predictor of contribution in the part with sanctions. There seem to be subject types when it comes to contribution choices.

When subjects punished, about 57.2% of the times they targeted the lowest contributor in their group. Interestingly enough, about 8.7% of the times they targeted the *highest* contributor in their group (perverse punishment). We scrutinized the subjects who targeted the highest contributor and why they did it. The pattern of perverse punishment was roughly stable over time. Revenge could explain some of it as the amount of punishment received is correlated with the frequency of perverse punishment. In particular, if the subject received no punishment in the previous period, she targeted the highest contributor with a frequency of 1.8% (N=4440). This frequency more than doubles if the subject received punishment (3.8%, N=5160). In particular, for a heavy punishment of eight or more points the frequency is 7.0% (N=644). Some perverse punishment may also be due to trembling hand. Only a minority of subjects though engaged in perverse punishment at least once (37.9% of subjects). They generally contributed less (4.1 tokens vs. 6.8) and requested more punishment than others (3.0 points per period vs. 1.7). Interestingly, their punishment was not exclusively perverse. In 46.3% of the cases they targeted the lowest contributor while in 35.1% of the cases they targeted the highest contributor.

The effectiveness of the consensual treatment in promoting cooperation could lie in its ability to censor perverse punishment. Most perverse punishment is requested by one person only (70%, all treatments) and hence was often not carried out. This contrast with punishment directed toward the lowest contributors, which was requested by one person only in 12.1% of the cases, hence highly likely to be carried out also in the consensual treatment. Table 4 reports estimates for requested versus actual punishment, which is in line with this interpretation. Punishment is generally lower when the target was the highest contributor in the group and this effect is stronger for actual than for requested punishment. This evidence concerns the relative

contribution of the target and complements the findings on absolute contribution showed in Table 2.

4. Conclusions

We study group cooperation in the provision of a public good under three peer punishment institutions, where agents have 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 of other experiments (Ostrom et al., 1992, Fehr and Gächter, 2002, Andreoni et al., 2003, Egas and Riedl, 2005), it also points to the significant impact of the specific punishment institution. There are three major conclusions.

First, the consensual institution of peer punishment performs remarkably better than the others ("consensus dividend"). This study has identified a specific set of rules that promotes a strong effect on group cooperation. Under a consensual institution, other-regarding preferences dominate the social interaction (Camerer and Fehr, 2006). When punishment toward a specific group member is requested by one agent only, the request to punish is ignored. Hence, there is actual punishment only when two or more agents requested it. Under a consensual institution, contributions and earnings are higher than when everyone has full discretionality on whom to punish. Without any external interference, this punishment rule aggregates individual norms within the group in a virtuous way that favors the emergence of the cooperative norm.

Second, we gained insights into the motivations that drive agents to punish. Changes in strategic incentives and information levels have surprisingly little effect in peer punishment behavior. Under the Sequential institution a subject about to punish a "target" individual knows how much other group members have already punished the individual. One would expect a

lighter punishment request if the target has already received a sanction and a heavier request otherwise. Instead, when the above information is provided the subject mostly ignores it, i.e. does not adjust her punishment request. We call this disregard for potentially useful information the "*coordination puzzle*." More work is needed on this point but it suggests that the punisher derives her utility from the act of punishing in itself and not from achieving, in conjunction with other punishers, a total amount of punishment that would discourage the free-rider. This interpretation casts doubts on the view that peer punishment is intentionally provided by subjects as a second-order public good (Ostrom et al., 1992, Sober and Wilson, 1998). According to this view subjects should care about the total punishment that another agent receives, and hence have no objections to others doing the "dirty job" of punishing. They should actually prefer it because it saves them the punishment cost. That may imply that when it comes to other-regarding attitudes, emotions alter the ability of people to behave strategically.

The third conclusion concerns the efficiency consequences of peer punishment. Peer punishment is not inherently efficiency-enhancing; it could damage group net earnings or boost them depending of what specific form of peer punishment is available in the social situation. We find that in two out of three treatments the ability to punish lowers net earnings. 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, Girard, 1977, p.16-22). Judicial systems are regulated forms of punishment that attempt to replace to some degree peer punishment in enforcing social norms. Our findings on the Consensual rule 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 research is needed to explore the behavioral foundations of punishment through legal systems.

References

Anderson, C. M., 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., Harbaugh, W., Vesterlund, L., 2003. The carrot or the stick: Rewards, punishment and cooperation. American Economic Review 93, 3, 893-902.

Bowles, S., Gintis, H., 2004. The evolution of strong reciprocity: Cooperation in heterogeneous populations. Theoretical Population Biology 65, 17-28.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the United States of America 100, 6, 3531-3535.

Camerer, C., Fehr, E., 2006. When does 'Economic Man' dominate social behavior? Science 311, 6, 47-52.

Carpenter, J., 2007. The demand for punishment. Journal of Economic Behavior and Organization 62, 522-542.

Casari, M., Plott, C.R., 2003. Decentralized management of common property resources: Experiments with centuries-old institutions. Journal of Economic Behavior and Organization 51, 217-247.

Casari, M., 2005. On the design of peer punishment experiments. Experimental Economics 8, 107-115.

Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? Experimental Economics 9, 265-79.

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

Decker, T., Stiehler, A., and Strobel, M., 2003. A comparison of punishment rules in repeated public good games. Journal of Conflict Resolution 47, 751-772.

Egas, M., Riedl, A., 2005. The economics of altruistic punishment and the demise of cooperation. Tinbergen Institute Discussion Papers 05-065/1.

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

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

Fischbacher, U., 2007. Z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171-78.

Gächter, S., Herrmann, B., 2006. The limits of self-governance in the presence of spite: Experimental evidence from urban and rural Russia. IZA Discussion Papers, no. 2236.

Girard, R., 1977. Violence and the Sacred. Baltimore: Johns Hopkins University Press.

Guererk, O., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312, 5770, 108-111.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. Costly punishment across human societies. Science 312, 1767-1770.

Houser, D., Xiao, E., McCabe, K., Smith, V., 2005. When punishment fails: Research on sanctions, intentions and non-cooperation. EconWPA, series Experimental, no. 0502001.

Lowie, R., 1970. Primitive Society. New York: W W Norton & Co..

Laurent, D.-B., Masclet, D., Noussair, C., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Economic Theory 33, 145-167.

Nikiforakis, N. S., 2008. Punishment and counter-punishment in public goods games: Can we really govern ourselves? Journal of Public Economics 92, 91-112.

Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: Self-governance is possible. American Political Science Review 86, 404-417.

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

Varian, H., 1994. Sequential contributions to public goods. Journal of Public Economics 53, 165-186.

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

Table 1: Individual contributions by session

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)

Notes: Sessions were conducted in 2003

Table 2: Punishment rates by contribution

Individual contribution	Baseline Avg. points	Sequential Avg. points	Consensual			
			Assigned (1)	Requested (2)	Difference (2) – (1)	% Censored [(2) – (1)] /(2)
0	4.55	3.62	4.84	5.13	0.28	5.5%
(0, 5]	1.98	2.26	1.36	1.77	0.41	23.1%
(5, 10]	1.30	1.62	1.62	2.11	0.49	23.2%
(10, 15]	0.49	1.24	1.34	2.03	0.69	34.1%
(15, 20]	0.56	1.42	0.25	0.93	0.68	73.2%
Total	2.41	2.54	1.70	2.20	0.50	22.8%

Notes: (1) Each individual contribution action is classified into one of five levels; then the average number of points of punishment received is computed (2,400 obs., i.e. 800 for each treatment). (2) The minimum number of observations in each cell of the Consensual columns is 113.

Table 3: 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%
(No. of observations)	(800)	(800)	(800)

Table 4: Determinants of punishment expenditure (punishment given)

	Baseline	**Sequential**			**Consensual**	
	all	all	*step 1*	*step 4*	All, actual punshment	*All, requests to punish*
Average subject's contribution without sanctions(periods 1-10)	-0.01 (0.07)	0.06 (0.06)	0.04 (0.05)	-0.01 (0.06)	0.06 (0.05)	0.05 (0.05)
Target was the *lowest* contributor in her group	2.22*** (0.29)	1.53*** (0.34)	1.84*** (0.43)	1.31*** (0.43)	2.42*** (0.24)	2.16*** (0.21)
Target was the *highest* contributor in her group	-1.29*** (0.24)	-0.74*** (0.22)	-1.16*** (0.38)	-0.02 (0.42)	-2.96*** (0.49)	-1.62*** (0.20)
Subject's contribution minus average group contribution	0.06* (0.04)	0.02 (0.04)	0.02 (0.05)	0.08 (0.05)	0.06** (0.03)	0.04* (0.02)
Average contribution of subject's other group members in previous period	0.01 (0.01)	0.02* (0.01)	0.02 (0.02)	0.01 (0.02)	0.005 (0.01)	0.004 (0.01)
Punishment received by other group members in previous period	0.06*** (0.02)	0.08*** (0.02)	0.05*** (0.02)	0.05** (0.02)	0.10*** (0.03)	0.10*** (0.02)
Punishment received by subject in the previous period	-0.006 (0.04)	0.02 (0.04)	0.02 (0.06)	0.06 (0.05)	-0.0009 (0.04)	0.02 (0.04)
Punishment received by subject in the second previous period	-0.01 (0.03)	-0.04 (0.03)	-0.11** (0.06)	-0.003 (0.05)	0.08** (0.03)	0.08*** (0.03)
The subject was alone in punishing the target in the period – dummy (a)	4.24*** (0.52)	4.24*** (0.36)	3.91*** (0.47)	5.04*** (0.62)		4.25*** (0.48)
Everyone punished the target in the period – dummy (b)	3.22*** (0.44)	3.00*** (0.37)	3.00*** (0.84)	3.42*** (0.49)	2.62*** (0.38)	2.54*** (0.36)
Step 1		0.14 (0.17)				
Step 4		-0.32** (0.14)				
Constant	-3.35*** (0.66)	-4.01*** (0.75)	-3.73*** (0.82)	-3.76*** (0.75)	-4.68*** (0.65)	-4.14*** (0.51)
Observations, no. subjects	2560, 80	2560, 80	640, 80	640, 80	2560, 80	2560, 80
Pseudo R-squared	0.175	0.153	0.177	0.158	0.206	0.220
Log likelihood	-2318	-2412	-650.4	-534.6	-1700	-2032
F-test: p-value for (a)=(b)	0.000	0.002	0.281	0.005	-	0.000

Note: (1) Tobits with individual random effects. (2) Dependent variable: request by subject i to punish subject k≠i; in every period there are four observations for each subject. (3) Session and period dummies were included in the

regression but are not reported. (4) Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. (5) If everyone in a group contributed the same amount there was neither a lowest nor a highest group contributor.

Table 5: Determinants of individual contribution to the public good under the punishment condition

	Baseline	**Sequential**	**Consensual**
Average subject's contribution without sanctions(periods 1-10)	0.47*** (0.18)	0.85*** (0.20)	0.36** (0.14)
Average contribution of subject's other group members in previous period	0.15*** (0.02)	0.12*** (0.03)	0.19*** (0.02)
Punishment received by other group members in previous period	0.20*** (0.04)	0.23*** (0.05)	0.35*** (0.07)
Punishment received by subject in the previous Period	-0.44*** (0.08)	-0.75*** (0.17)	-0.65*** (0.12)
Punishment received by subject in the second previous period	-0.13** (0.06)	-0.43** (0.17)	-0.26*** (0.10)
Punishment expenditure of the subject in the previous Period	-0.16* (0.08)	-0.17 (0.10)	-0.28** (0.13)
In the previous period the *lowest* contributor in her group was punished	0.36 (0.72)	3.23** (1.38)	0.44 (0.60)
In the previous period the *highest* contributor in her group was punished	-1.47*** (0.41)	-1.12** (0.55)	-2.64* (1.47)
Constant	-1.63 (1.12)	-4.52** (1.87)	3.64* (1.91)
Observations, No.subjects	640, 80	640, 80	640, 80
Pseudo R-squared	0.118	0.101	0.155
Log likelihood	-1494	-1378	-1619

Note: (1) Tobits with individual random effects. (2) Session and period dummies were included in the regression but are not reported. (3) If everyone in a group contributed the same amount there was neither a lowest nor a highest group contributor. (4) Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Figure 1: Contribution to the public good over time