



Working Paper Series  
Department of Economics  
University of Verona

# Legitimate Punishment, Feedback, and the Enforcement of Cooperation

Marco Faillo, Daniela Grieco, Luca Zarri

WP Number: 16

November 2010

ISSN: 2036-2919 (paper), 2036-4679 (online)

# Legitimate Punishment, Feedback, and the Enforcement of Cooperation

Marco Faillo<sup>a</sup>

Daniela Grieco<sup>b</sup>

Luca Zarri<sup>c</sup>

## *Abstract*

In real life, punishment is often implemented only insofar as punishers are entitled to punish and punishees deserve to be punished. We provide an experimental test for this principle of legitimacy in the framework of a public goods game, by comparing it with a classic (unrestricted) punishment institution. A significant advantage of our institution is that it rules out antisocial punishment, a phenomenon which recent studies document to play a key role in undermining the scope for self-governance. Our findings show that, despite the lack of additional monetary incentives to cooperate, the introduction of legitimate punishment leads to substantial efficiency gains, in terms of both cooperation and earnings. Therefore, in businesses and other organizations, this device could successfully deal with the principal-agent problem, with the principal delegating a task to a team of agents. Further, we interestingly find that removing the information over high contributors' choices only leads to a dramatic decline in cooperation rates and earnings. This result implies that providing feedback over virtuous behavior is necessary to make an institution based on legitimate punishment effective.

**JEL Classification:** C73;C91; D02 ; D63.

**Keywords:** Experimental Economics; Public Goods Games; Costly Punishment; Cooperation; Legitimacy; Feedback.

<sup>a</sup>[marco.faillo@unitn.it](mailto:marco.faillo@unitn.it), Department of Economics, University of Trento.

<sup>b</sup>[daniela.grieco@univr.it](mailto:daniela.grieco@univr.it), Department of Economics, University of Verona.

<sup>c</sup>[luca.zarri@univr.it](mailto:luca.zarri@univr.it), Department of Economics, University of Verona.

## 1. Introduction

In naturally occurring environments, punishment is a widespread phenomenon. A typical feature of sanctioning mechanisms, both within formal and informal institutions, is that their usage is far from being arbitrary and unrestricted. Everyday life abounds in examples where specific requirements have to be met in order for a person or an institution to be viewed as a potential punisher as well as a potential punishee. In many countries, you need to have a clear criminal record to apply for jobs such as police officer or judge, where you will need to sanction others on a daily basis. Even the authority of figures like school teachers and parents might be at risk if they misbehave or if their behaviour is not in line with what they are trying to teach to their students and kids. Elected politicians will act as lawmakers, but if, say, a member of parliament known for his tough anti-drugs or anti-prostitution campaigns gets caught at a party with cocaine or escorts, in many countries the media will easily induce him to resign. At the international level, in the current political debate on the hot topic of nuclear weapons development, a forcefully repeated claim is that while democratic countries (e.g. Israel) are entitled to produce nuclear weapons, non-democratic regimes (e.g. Iran and North Korea) are not. What these otherwise distant situations where punishment is at work have in common is an underlying principle of *legitimacy*: only *some* people or institutions have the right to sanction ('entitlement') and *not everyone* deserves to be sanctioned ('desert'). In modern societies, punishment is usually viewed as socially and ethically acceptable only insofar as a principle of legitimacy holds. Centuries of normative argument in applied ethics, philosophy of law and political philosophy (with classical contributions from prominent thinkers such as John Stuart Mill and, more recently, John Rawls, Jurgen Habermas and Ronald Dworkin, among many others) have convincingly made clear that in a liberal democracy punishment needs to be legitimate, in order to be theoretically justified<sup>1</sup>. Also in his influential classical paper on crime and punishment, Becker (1968) takes for granted that punishment must be legitimate in order to be allowed. In this article, we investigate this legitimacy-punishment nexus experimentally within a public goods game framework and find that legitimate punishment turns out to be an effective institution in both enforcing cooperation and raising individual earnings. We also focus on the role of information and show that restrictions on punishment activity are effective only when feedback over how the most deserving individuals do actually behave is provided. On the whole, then, our results interestingly suggest that it is the *interaction* between the legitimate nature of the sanctioning institution at work and the amount of information over peers' contribution behavior provided to the

---

<sup>1</sup> On philosophical grounds it can be plausibly maintained that the very existence of the modern state itself rests upon a fundamental legitimacy argument: in a democracy, citizens delegate the power to the state and, due to its being the legitimate representative of the people, the government has access to coercive power. Within their geographical boundaries, states are sovereign and allowed to sanction citizens adopting wrongful behavior right because society as a whole conferred to them the legitimacy to do so.

subjects that plays a critical role in determining final contribution and earning levels. The remainder of the paper is structured as follows: section 2 briefly reviews the related literature; section 3 illustrates the experimental design; section 4 reports our main results and section 5 discusses our findings and concludes the paper.

## 2. Related literature

In a public goods game or voluntary contribution mechanism (*VCM*), there is a group of subjects who, as the game starts, receive an individual monetary endowment, from which they may contribute any amount to a public good that returns a payoff to each of them. The structure of monetary payoffs in the *VCM* makes it a classical ‘social dilemma’, as each agent has a dominant strategy to free ride, while, in contrast, at the social optimum each individual contributes his entire endowment. Therefore, the straightforward prediction based on so called *Homo Oeconomicus* is that everyone should free ride, both in the one-shot and in the finitely repeated game. However, in the finitely repeated version of this game, the following pattern typically occurs: initially, average contributions are relatively high, whereas, as the game unfolds, they gradually decline and cooperation converges to a near-negligible level (Ledyard, 1995).

In the last years, an increasing number of *VCM* experiments have been investigating the role that institutions can play in the enforcement of cooperation. While a strand of experimental research deals with endogenously formed institutions (see e.g. Gürer et al., Kosfeld et al., 2009 and Sutter et al., 2010), a second strand encompasses exogenously imposed institutions. Within the latter research area, some studies focused on centralized mechanisms (see Chen and Plott, 1996; Falkinger et al., 2000; Andreoni, 1993; and Chan et al., 2002), whereas others explored decentralized institutions (Ostrom et al., 1992; Fehr and Gächter, 2000; 2002; Casari and Plott, 2003; Rand et al., 2009; Fudenberg and Pathak, 2010). In their pathbreaking study, Fehr and Gächter (2000; 2002) demonstrate that while in non-punishment treatments (*VCM* without punishment opportunities) cooperation rates indeed tend to fall over time (round after round), this ‘decay phenomenon’ does not occur insofar as players are allowed to incur a cost to decrease others’ monetary payoffs (*VCM* with punishment opportunities). The presence of punishment opportunities turned out to make the difference and made cooperation sustainable over time<sup>2</sup>. Insofar as we suppose that in the laboratory subjects act selfishly in order to systematically maximize their monetary gains, costly punishment is a puzzle. In other words, this behavioral assumption predicts that, in a finitely repeated *VCM* with punishment options, subjects *will not* use

---

<sup>2</sup> Analogously, the introduction of explicit punishment and/or rewarding opportunities significantly affects subjects’ behavioral choices in the experimental games studied by Falk and Kosfeld (2006) and Fehr and Rockenbach (2003).

such options, due to the net monetary costs associated with their usage<sup>3</sup>. By contrast, peer punishment of free riders turned out to be a widespread phenomenon both in the field and in the lab, where it occurred both with anonymous random matching (Fehr and Gächter, 2000; 2002; Egas and Riedl, 2008; Anderson and Putterman, 2006; Rockenbach and Milinski, 2006) and with fixed groups playing a finite number of times (Yamagishi, 1986; Fehr and Gächter, 2000; Page et al., 2005; Nikiforakis and Normann, 2008). Experimentally, it has been shown to represent a powerful enforcement device, through which it is possible to induce and successfully sustain cooperation in social dilemmas.

Like these studies, in this contribution we focus on a decentralized mechanism based on exogenously determined sanctioning opportunities. However, unlike the papers cited above, but consistently with the considerations developed in the introductory section, we suitably restrict access to punishment options: our institutional arrangement is based on ‘legitimate punishment’ in the sense that it prescribes that only relatively high contributors can sanction and only relatively low contributors can be sanctioned. While some of the previously cited papers focus on institutions which derive their legitimacy from a *process* of endogenous choice, we analyze an enforcement device which is exogenously imposed but at the same time legitimate due to its inner, structural features, i.e. due to its conditioning the possibility to punish on the adoption of cooperative behavior in the first place. By focusing on restricted punishment, we depart from most of the existing experimental literature on punishment, as lab studies on sanctioning mechanisms have mainly focused on *unrestricted* punishment. In a public goods game environment, unrestricted punishment seems to work extremely well, under certain conditions. Fehr and Gächter’s (2000) well-known findings represent a very important ‘spontaneous order’ result: subjects are willing to sanction others even if this is costly and such an institution is effective in enhancing cooperation and preventing the unpleasant ‘decay phenomenon’ occurring when punishment options are unavailable. However, recent work convincingly reveals that there is also a ‘dark side’ of unrestricted punishment. In particular, the following important four drawbacks have been identified in the last years: (1) the quantitative relevance of antisocial punishment; (2) the lack of robustness to institutional changes; (3) the risk of motivation crowding-out and (4) the low level of average earnings. Let us shortly illustrate each of these downsides of unrestricted punishment. First, unrestricted punishment in many cases significantly undermines the scope for self-governance, as, since everyone is free to punish everyone else, sanctioning may take the form of misdirected, ‘antisocial’ punishment – that is, low contributors punishing high contributors. Recent evidence indicates that antisocial punishment substantially reduces contribution rates (Cyniabuguma et al.,

---

<sup>3</sup> Sethi and Somanathan (1996) observe, on the basis of the case studies cited in their work, that punishments such as social disapproval and physical damage are costly not only for the punishee, but also for the punisher.

2006) – especially if it is targeted at outgroup members when competition between groups is created (Goette et al., 2010) or it occurs within less industrialized societies (Herrmann et al., 2008) –, to the point that cooperation in the presence of punishment can be even *lower* than in its absence (Gächter and Herrmann, 2010). As Gächter and Herrmann (2010) correctly point out, “Punishment of cooperators has been largely neglected in previous research on social preferences because it was negligible compared to the punishment of free riders. Our results show that this neglect is not warranted because punishment of cooperators can be very significant in some subject pools”. Second, when multiple stages of punishment are allowed, counterpunishment and feuds are likely to be triggered, limiting, once again, successful self-governance and leading, eventually, to a demise of cooperation (Nikiforakis, 2008 and Nikiforakis and Engelmann, forthcoming). Since the opportunity to retaliate punishments exists in many real-life decentralized interactions (Nikiforakis, 2008), these negative results show that unrestricted punishment is not robust, as an effective cooperation enforcement device, to minimal institutional changes. Third, a further problem is that since this form of punishment exclusively relies on deterrence, that is on *extrinsic* motives to cooperate, the risk is either not to elicit people’s *intrinsic* motivations to comply or even to crowd them out, especially when incentives are weak<sup>4</sup>. Fourth, another crucial point to be made is that solving the free rider problem and achieving a significant level of cooperation is only one part of the problem as a whole. In particular, recent papers indicate that, even in the presence of a single stage of sanctioning, the success of unrestricted punishment in enforcing cooperation comes at a substantial cost. Botelho et al. (2005) analyzed Fehr and Gächter’s (2000; 2002) data and find *lower earnings* when punishment was allowed than under no punishment (see also Cyniabuguma et al., 2006 for similar results on this and other sanctioning experiments). The same occurs to average payoffs in 13 out of 16 participant pools of Herrmann et al.’s (2008) experiment on antisocial punishment cited above. Similarly, Dreber et al. (2008) show that when in a repeated prisoner’s dilemma players can choose between cooperation, defection and costly punishment, average group payoffs are not higher than when the punishment option is not available. Further, since punishing is costly not only for the punishees but also for the punishers, the ‘winners’ (i.e. those who get the highest earnings) in their experiment are the individuals who abstain from punishing. This evidence indicates that unrestricted punishment is a double-edged sword (Goette et al., 2010), as it raises cooperation levels but, unless we consider a significantly longer time horizon (Gächter et al., 2008), leads to average earnings which are lower than in the absence of sanctioning opportunities. From an economic perspective, this is a serious shortcoming of unrestricted punishment, showing that this

---

<sup>4</sup> Fehr and Rockenbach (2003) provide experimental evidence that sanctions underlying selfish or greedy intentions – unlike sanctions perceived as fair – produce extremely negative effects on cooperation.

form of sanctioning risks to determine efficiency losses and, therefore, to turn into a wasteful activity for those societies or organizations that adopted it.

On the whole, these arguments strongly question the belief that individuals are able to successfully govern themselves through punishment (Nikiforakis, 2008). A natural solution seems then to rely on an exogenous, central authority, by assigning to a Hobbesian Leviathan the power to sanction non-cooperators. However, monitoring the individuals can be extremely costly and an important implication to be drawn from this paper that this needs not be the case, as solving the above drawback does not automatically imply passing from decentralized to centralized punishment. Decentralized punishment can be successful. The key condition for this to occur is that it needs to be suitably restricted, along the lines described above. In particular, one reason why we decided to investigate legitimate punishment is that we expected such an institution not to suffer from the limitations which the recent studies cited above have found with regard to unrestricted punishment. While with unrestricted punishment various forms of undesirable behaviors often occur and lead, over time, to a breakdown of cooperation, a punishment institution based on a principle of legitimacy rules out detrimental forms of sanctioning – such as (first-order) antisocial punishment and, when multiple stages of punishment are allowed, feuds, counterpunishment and higher-order perverse punishment – by construction. On positive grounds, it was also plausible to expect that such a principle of legitimacy may elicit people's intrinsic motivation to contribute and punish low contributors. Moreover, in the light of these considerations, we also wondered whether under legitimate punishment aggregate earnings could be higher than under unrestricted punishment<sup>5</sup>.

It is reasonable to believe that in this context the impact of punishment on cooperation could also depend on the amount of information about others' behavior, a variable which has been almost completely neglected in the punishment literature<sup>6</sup>, and which could have a significant influence on the perception of the legitimacy of the sanction. We believe that within an environment in which the right to punish is awarded on a meritocratic basis, feedback over how the most virtuous members of the group behave might play an important two-fold role in promoting cooperation. First, when this information is provided, a member who has been punished is not only aware of the fact that her contribution to the public good is lower than the contribution of the member who has punished her, but she also knows the exact level of contributions of those who have gained the right to punish. In this sense, the provision of information on the most virtuous members' choices contributes to shed

---

<sup>5</sup> Related papers where punishment is not unrestricted include Ertan et al. (2009), Xiao and Kunreuther (2010) and Casari and Luini (2009). In the latter, the authors allow punishment only insofar as it is requested by a coalition of at least two subjects.

<sup>6</sup> For exceptions, see Nikiforakis (2010), Xiao and Houser (2010), and Grechenig et al. (2010). As Nikiforakis (2010) points out, institutional details such as the format in which feedback about the actions of others is given can affect the efficacy of peer punishment in promoting cooperation.

light on the degree of legitimization of the punishment activity. Second, this kind of feedback could also serve a pure cognitive function, as an individual who knows how the virtuous members of her group behave also knows what she must do to avoid punishment in the next future and what is the level of contributions expected by the other group members.

In our study, we address this problem by comparing the case in which subjects have information on every other member's contributions with the case in which each member is informed only about the average contribution of her group and on the contribution of the members who have contributed less than herself. In the latter case, members whose contribution is not the highest do not know what is the highest level of contribution in their group.

### 3. Experimental setup

In our sanctioning institution, some key restrictions are exogenously imposed with regard to both *who* is allowed to punish and *whom* punishers can punish<sup>7</sup>. These assumptions are in line with what happens within several naturally occurring environments like the ones recalled in the introduction, where it is often the case that the social acceptance of punishment is conditional on (i) the punisher being entitled to punish (*entitlement*) and (ii) the punishee being a wrongdoer and, therefore, deserving to be punished (*desert*). When the two requirements of entitlement and desert are met, we say that punishment is legitimate (i.e. a principle of legitimacy holds).

Since we investigate a finitely repeated *VCM* with punishment options, a two-stage game gets played in every period: at stage 1, players simultaneously choose how much to contribute to the public good (contribution stage) and at stage 2 they have access to punishment options (punishment stage). However, the principle of legitimacy requires that a single individual acts as a 'high contributor' at stage 1 in order to earn the right to be a punisher at stage 2<sup>8</sup>. More specifically, we assume that a subject is entitled to punish another subject at stage 2 only if her contribution at stage

---

<sup>7</sup> Therefore, our design also differs from recent experimental *VCM* protocols where norms prescribing who can punish and/or who can be punished emerge endogenously within a group (see e.g. Casari and Luini, 2009; Kosfeld et al., 2009). Casari and Plott (2003) is an example of an experimental paper where, like in the present setup, 'virtuous' restrictions on punishment are exogenously imposed. Xiao and Houser (2010) assume that when a round is monitored, then that round's lowest contributor will incur a small sanction. However, they assume that punishment is not peer-to-peer but exogenous, that is under the experimenters' control.

<sup>8</sup> As far as immediate monetary consequences of subjects' sanctioning decisions are concerned, it is worth noting that while in Casari and Plott (2003) the subjects who find and sanction free riders are monetarily rewarded, in our design legitimacy, by allowing cooperators to have access to punishment options, only makes them entitled to costly punish wrongdoers. Xiao and Kunreuther (2010) compare deterministic vs. stochastic punishment in the framework of a prisoner's dilemma game and, in two out of six treatments, introduce a rule such that, like in the present paper, only cooperators are allowed to punish non-cooperators. However, studying the impact of restricted punishment in a two-player game like the prisoner's dilemma, where each player always knows who punished whom, significantly differs from investigating the effectiveness of legitimacy in a multi-player environment like the *VCM*.

1 has been *higher* than the contribution of the peer she wants to punish<sup>9</sup>. As a consequence, high contributors are (partially) immune from punishment, in the sense that they cannot be sanctioned by players who contributed less than them. This rules out antisocial punishment (Herrmann et al., 2008). Like in a standard, finitely repeated *VCM*, insofar as all the subjects are supposed to be driven by material self-interest only and this information is common knowledge, the unique subgame perfect equilibrium is for all agents to *never punish* and *never contribute*.

### 3.1 Procedure

A total of 168 subjects participated voluntarily in the experiment at the CEEL Lab of the University of Trento. A total of 9 sessions were conducted, between December 2009 and November 2010. Six sessions had 20 participants and the other three had 16 participants. The experiment was programmed by using the z-tree platform (Fischbacher, 2007). The subjects, were undergraduate students (64.3% from Economics, 49.5 % females, 80.3 % Italian). No individual participated in more than one session. In each session, the participants were paid a 5 euro show up fee, plus their earnings from the experiment. The average payment per participant was 15.70 euros (including the show-up fee) and the sessions averaged approximately 1 hour and 30 minutes. At the beginning of each session, participants were welcomed and asked to draw lots, so that they were randomly assigned to terminals. Once all of them were seated, the instructions<sup>10</sup> were handed to them in written form before being read aloud by the experimenter. We took great care to ensure that the participants understood both the rules of the game and the incentives. They had to answer several control questions and we did not proceed with the actual experiment until all participants had answered all questions correctly.

In each session, there are 20 periods of interaction that proceed under identical rules. The participants in a session were randomly assigned to groups of size four, so that they did not know the identities of the other members of their group. Like other experimental studies (see e.g. Cinyabuguma et al., 2006; Denant-Boemont et al., 2007), we used a partner protocol that kept the composition of each group constant over rounds, so that, at the end of each period, individuals remained in the same group. We did this as repeated interaction is a typical feature of several

---

<sup>9</sup> This implementation of the principle of legitimacy differs from the prevailing form of restricted punishment endogenously emerging in Ertan et al. (2008). In their public goods game experiment, subjects vote on whether to allow sanctioning of group members whose contributions are (a) below-average, (b) above-average and (c) equal to the average: it turns out that eventually the majority of groups opt for prohibiting punishment of higher-than-average contributors. Noussair and Tan (2009) investigate whether this ability of a voting process to converge to the optimal institutional structure is robust to a specific change in the environment, that is the existence of heterogeneity in the value to the group of subjects' contributions. While their results extend the findings of Ertan et al. (2008), the two authors also find that agents fail to converge (through voting) to the efficient punishment regime.

<sup>10</sup> A translation of the instruction sheet is provided in Appendix A. Original instructions were written in Italian. They are available upon request from the authors.

naturally occurring environments (e.g., businesses or collectives) where punishment occurs (Xiao and Houser, 2010). However, individuals' labels were reassigned on a random basis in each period. For example, the same player could be designated as player 45 in period  $t$ , as player 6 in period  $t + 1$ , and as player 38 in period  $t + 2$ . Therefore, our partner protocol was also characterized by anonymity of the components of the group and change of participants' labels across rounds<sup>11</sup>. The design and the parametric structure of the experiment are based on those of Fehr and Gächter (2000).

### 3.2. Treatments

We implemented three experimental treatments: a baseline, unrestricted punishment and full information (Baseline) treatment, a restricted punishment with full information (Full R.) treatment and a restricted punishment with partial information (Partial R.) treatment.

There were 3 sessions (20 subjects in two sessions and 16 in the other) for the Baseline, 3 sessions (with 20 subjects in two sessions and 16 in the other) for the Full R. and 3 sessions (with 20 subjects in two sessions and 16 in the other) for the Partial R. For each treatment, in each session the subjects were divided in groups of  $N=4$  (as in standard *VCM* experiments) subjects, who played a two-stage finitely repeated public goods game with punishment options for  $T=20$  periods. Participants were aware of the number of rounds they were going to play and of the number of stages: information on the following stages allows to evaluate the effect of the threat of being punished in stage 2 and on contribution decisions in stage 1.

Overall, the three treatments differ along two dimensions (see Table 1): *behavioral restrictions* and *feedback about others' contribution levels* in the group.

[TABLE 1 ]

In the Baseline treatment, punishment is unrestricted and subjects are provided with full information, that is there is feedback about all their group co-players' individual contributions. This is a replication of the standard *VCM* with punishment (Fehr and Gächter, 2000), where everyone can freely punish everyone else in the group. The other two treatments are both based on legitimacy (i.e. entitlement and desert): both in the Full R. and the Partial R. treatment, a subject

---

<sup>11</sup> Although a stranger protocol with random re-matching allows ruling out strategic punishment and reputation motives altogether, a partner protocol seems to work as well as a stranger protocol. Nikiforakis (2008), based on Botelho et al. (2005), addresses this issue by comparing results from a stranger protocol and a partner protocol and finds that differences in punishment decisions are not significant (whereas differences in punishment levels are). Fehr and Gächter (2000) find differences in outcomes between the partner and the stranger protocol in their *VCM* experiment.

is entitled to sanction another subject in stage 2 only if her contribution at stage 1 has been *higher* than the contribution of the peer she wants to punish. The difference between the two treatments regards the feedback that subjects receive at the end of stage 1, in each period: while in Full R. subjects are informed about the full vector of others' contributions (like in the Baseline), in Partial R. subjects are informed only about the *average* contribution level and the specific contribution levels of their group co-players who contributed *less* than them. Therefore, no specific information about more virtuous peers is provided to them in this treatment.

### 3.2.1. Legitimacy-based treatments

While our Baseline treatment is based on the standard *VC*M with punishment options (Fehr and Gächter, 2000), our two legitimacy-based treatments (Full R. and Partial R.) share the following features. In stage 1, at the beginning of each period each participant receives a fixed amount  $e = 20$  of tokens<sup>12</sup>. Each participant  $i$  has to decide whether she wants to invest into a public project or not an amount  $g_i \leq e$ . Decisions are made simultaneously and with no information about peers' choices. At the end of stage 1, each participant is informed about her current earnings, which consist of two elements:

- a. The amount of her initial 20 tokens that she has kept for herself (i.e. 20 tokens – Her Contribution to the project);
- b. Her income from the project. The income to her is equal to 40% of the total of the four individual contributions to the project.

Therefore, her earnings at the end of stage 1 are calculated by the computer in the following way:

Each participant's earnings after stage 1 = (20 – her contribution to the project) + 40%\*(total group contribution to the project).

---

<sup>12</sup> 1 token = 0,02 euro.

Participants know that they can go on with stage 2 in the experiment only if they contribute more than their peers, that is, as we explained above, only if they are entitled to do so<sup>13</sup>. Specifically, player  $i$  will be entitled to sanction player  $j$  in stage 2 only if  $g_i > g_j$ . In stage 2, subjects are given the opportunity to simultaneously punish those who contributed less than them by assigning a certain amount of points. This implies that the highest contributor in a group is fully immune from punishment. Potential punishers might decide to assign up to 10 points to each punishee: point assignment is costly and costs are charged according to a standard cost function as in Fehr and Gächter's (2002) (Table 2).

[TABLE 2]

Each point that a subject receives reduces her earnings at stage 1 by 10%.

Each participant's earnings at the end of stage 2 are calculated by the computer in the following way:

Each participant's earnings after stage 2 = earnings at the end of stage 1 - cost of points she assigned at stage 2 - 10%\* number of points received\*earnings at the end of stage 1

## 4. Results

### 4.1. Contribution levels

Figure 1 displays the time pattern of individual contributions by period, averaged across groups, in the three treatments.

[FIGURE 1]

---

<sup>13</sup> It is important to make clear that we never used loaded terms such as 'legitimacy', 'entitlement', 'desert', 'punishment', 'free riding' and 'immunity' during the experiment.

In all the treatments contribution levels do not decline over time.

*Result 1. Punishment prevents the decline of cooperation over time in all the treatments.*

[TABLE 3]

Besides this well-known general positive effect of punishment, our data show (Table 3) that, given the same type of restrictions on the punishment activity, subjects who are informed about the contributions of all the other members of their group (Full R. treatment) contribute significantly more than subjects who are informed only about the average contribution of their group and on less virtuous peers' contributions (Partial R. treatment) (Wilcoxon Rank-sum Test with group averages as observation:  $z=2.43$ ; p-value: 0.014). At the same time, given the same level of information, contributions in the Full R. treatment are on average significantly higher than contributions in the baseline treatment ( $z=2.61$ ; p-value: 0.08). The introduction of restrictions on the punishment activity has a positive effect on the level of contributions. These differences characterize also the distribution of contributions in the final period of the game (Figure 2). Result 2 follows.

[FIGURE 2]

*Result 2. The introduction of restrictions increases the level of cooperation only if detailed information on the contribution levels within the group is provided<sup>14</sup>.*

This result is supported by the regression analysis<sup>15</sup> reported in Table 4, which takes into account the effect of a set of control variables and sheds further light on the role of restrictions and information in shaping the contribution levels.

[TABLE 4]

Besides the treatment effect, contributions in each period are positively (and significantly) affected by the average contribution in the group in the first period (variable AV\_first). Therefore, each

---

<sup>14</sup> The levels of contribution observed in the Partial.R and in the Baseline are not significantly different (Wilcoxon Rank-sum Test with:  $z=-0.046$ ; p-value: 0.96). Note however that a direct comparison between the Baseline and the Partial.R treatments is not particularly useful, since Partial.R differs from the Baseline both for the presence of restrictions and for the quantity of information provided to the subjects.

<sup>15</sup> All the estimations have been carried out with STATA 11.

group's behavior in the first period represents a key determinant of subsequent contribution choices in the group: cooperation seems to be sustained also by idiosyncratic features of the specific group. Higher contribution in the Full R. treatment also results in a higher level of efficiency (figure 3). Taking group average earnings as independent observations, we observe that average earnings in the Full R. treatment are significantly higher than average earnings both in the Baseline (Wilcoxon Rank-sum Test:  $z=2.52$ ; p-value: 0.011) and in the Partial R. treatment (Wilcoxon Rank-sum Test:  $z=2.89$ ; p-value: 0.003)<sup>16</sup>.

[FIGURE 3]

*Result 3. Average earnings are significantly higher when punishment activity is restricted and subjects have information on the contributions of all the other members of their group.*

#### **4.2. Punishment behavior**

As Result 2 shows, the introduction of restrictions in the aim of preventing the assignment of punishment points to virtuous subjects results in higher contribution levels. In order to account for this evidence we shall give a closer look at the punishment activity in the three treatments and assess the impact of antisocial punishment in the Baseline treatment.

[FIGURE 4]

With regard to the distribution of punishment points, in all the treatments we observe the typical decreasing pattern, which is faster in the Full R. treatment (Figure 4). The difference between the average quantity of points assigned in the three treatments is not statistically significant (Table 5) (Wilcoxon rank sum Full R. vs Partial R.:  $z=-1.19$ ; p-value= 0.23; Wilcoxon rank sum Full R. vs Baseline:  $z=-0.87$ ; p-value= 0.38).

[TABLE 5]

However, it is worth noting that in the Baseline treatment a non-negligible percentage of punishment points are assigned to virtuous subjects. Table 6 reports the absolute quantities (column 2) and the percentage (column 3) of punishment points assigned in the Baseline treatment by a subject  $i$  to a subjects  $j$  when the contribution of  $i$  is smaller than the contribution of  $j$ . We define

---

<sup>16</sup> The result is robust to controls for average contribution in the first period, quantity of assigned points, quantity of received point, gender, age, nationality, major and number of past experiments.

this type of behavior as “weak antisocial punishment”, as distinguished from “strong antisocial punishment”. The latter is observed when  $i$  punishes another subject  $j$  whose contribution is greater than both the contribution of  $i$  and the average contribution of the group (columns 4 and 5). In our sample 19.5% of the overall punishment activity (number of punishment points assigned in all periods) can be classified as weak-antisocial, while 12.2% is strongly antisocial. On average 14.4% of group’s punishment points assigned is weakly antisocial and 9% is strongly antisocial.

[TABLE 6]

[FIGURE 5]

The presence of a strong form of punishment of virtuous subjects (strong antisocial punishment) in the Baseline treatment emerges also in Figure 5, which displays the relation between the distance from the average contribution of the group and the average quantity of points received. In the Baseline treatment, in some cases strong positive deviations are still punished. This evidence is supported by the results of the following regression analysis (results in table 7):

$$\text{punishment points received}_{igt} = \beta_0 + \beta_1 \text{pos\_dist\_av}_{igt} + \beta_2 \text{neg\_dist\_av}_{igt} \text{ (Eq. 1)}$$

where  $\text{pos\_dist\_av}_{igt}$  is the positive distance from the group’s average contribution, i.e. the difference between the subject’s contribution and the group average contribution; this variable is equal to zero when the subject’s contribution is below the average. The variable  $\text{neg\_dist\_av}_{igt}$  is the absolute negative distance from the average of the groups, i.e. the absolute value of the difference between the group’s average contribution and the subject’s contribution; it is equal to zero when the subject’s contribution is above the average.

[TABLE 7 ]

While in all the treatments the quantity of punishment points received decreases as the negative distance from the average increases, positive distance from the average has a significant effect on the quantity of points received only in the two treatments with restrictions.

*Result 4. When the punishment activity is unrestricted, a non-negligible percentage of points are assigned also to subjects who contribute more than the punisher (weak antisocial punishment) and in some cases also to the most virtuous subjects (strong antisocial punishment).*

Result 4 is compatible with the higher level of contributions observed in the Full R. treatment, where both weak antisocial and strong antisocial punishment are ruled out.

#### 4.3. Determinants of changes in individual contribution levels

As we have shown in the previous subsections, the three treatments are significantly different in terms of contributions levels, but not in terms of punishment points assigned. Hence, an analysis of the effects of punishment in altering contribution levels is needed. In particular, we test if high contributors and low contributors' reactions to punishment are different. Having observed that a non-negligible share of punishment activity in the treatment without restrictions (Baseline) can be classified as antisocial, we shall investigate whether this punishment has also a perverse effect on the contribution level of the most virtuous members of the group<sup>17</sup> - i.e. whether it weakens their willingness to cooperate. In order to do this, the following equation is estimated for each treatment, distinguishing between subjects whose contribution is below the average contribution of the group and subjects whose contribution is not below the average of the group:

$$contribution_{igt} - contribution_{igt-1} = \beta_0 + \beta_1 received\_punishment_{igt-1} + \beta_2 dist\_av_{igt-1} \text{ (Eq. 2)}$$

where *received\_punishment<sub>igt-1</sub>* represents the number of punishment points that the subject has received in the previous period, whereas *dist\_av<sub>igt-1</sub>* is the distance between the subject's contribution and the average contribution in the group in the previous period. Results of the estimation are reported in Table 8, which shows a regression to the mean in all the treatment observed also by Denant-Boemont (2007): the higher the distance from the average in the previous period, the higher is the absolute increase of the contribution level in the current period.

With regard to the effect of punishment, we observe a positive and significant effect on low contributors' change in levels of contribution in the two treatments with restrictions (Full R. and Partial R.). The same effect is not observed for low contributors of the Baseline. Moving to high contributors, in the Baseline treatment we observe a negative reaction to punishment. The opposite effect is observed in the treatment with full information and restriction (Full.R), while high contributors in the Partial.R do not show any significant change in the level of contribution as a consequence of punishment. This evidence confirms the presence of a significant perverse effect of antisocial punishment that can explain the low level of contributions observed in the Baseline. The

---

<sup>17</sup> For a detailed analysis of this effect see Ones and Putterman (2007).

introduction of restrictions prevents this effect because high contributors know that punishment points come from the most virtuous members of their group.

[TABLE 8]

*Result 5. In all the treatments, regardless of the presence of restrictions, the increase in contribution levels is stronger the higher the distance from the average in the previous period.*

*Result 6a. Punishment has a positive effect on low contributors' willingness to cooperate only in the presence of restrictions.*

*Result 6b. Punishment has a negative effect on high contributors' willingness to cooperate in the Baseline treatment, while it has a positive effect in the case of high contributors in the treatment with full information and restrictions.*

Finally, in the aim of exploring the role of information about others' behavior in shaping contribution reactions to punishment, we estimate the following equation for each treatment, by considering only the subsample of subjects whose contribution in the previous period was not the highest:

$$contribution_{igt} - contribution_{igt-1} = \beta_0 + \beta_1 received\_punishment_{igt-1} + \beta_2 dist\_av_{igt-1} + \beta_3 dist\_highest_{igt-1} (Eq. 3)$$

Where  $dist\_highest_{igt-1}$  is the distance between subject's own contribution in period t-1 and the highest contribution in the group in period t-1. We run two separate estimations, by distinguishing between subjects whose contribution level in the previous period is *below* the average of the group and subjects whose contribution level in the previous period is *above* the average of the group (Table 8). In the case of subjects with contribution levels below the average, information on the most virtuous peers does not affect *per se* the increase in contributions in the Full R. treatment, i.e. in the treatment where the full vector of peers' contribution is available, antisocial punishment is ruled out and subjects have the possibility to use virtuous peers' behavior as a reference point. The evidence on the Baseline is particularly interesting. In this case, for subjects who contribute below the average, the distance from the virtuous subjects has a significant and negative effect on the change in contribution levels: the lower the subject's contribution at t-1 with respect to the highest contribution of her group at t-1, the lower the increase in her contribution moving from t-1 to t. These subjects seem to use the information on most virtuous peers to infer the extent at which they

can behave as free riders: the more altruist their peers are, the more profitable the choice of behaving as a free-rider.

With regard to subjects who contribute above the average, in the baseline treatment the information about the highest level of contribution in the group does not exert neither a positive nor a negative significant effect on the increase in contributions. In the Full R. treatment, the highest level of cooperation is taken as a reference point.

[TABLE 9 ]

*Result 7. In the full information treatment with restrictions, the highest contribution level in the group is used as a reference point only by subjects who contribute above the average of the group. In the full information treatment without restrictions, subjects who contribute below the average of the group use the information on the most virtuous members to infer the potential gain from free riding.*

#### **4.4. An extension: Average-dependent legitimacy**

The legitimacy-based treatments (Partial R. and Full R.) described above present a definition of legitimacy that is fully exogenous, in the sense that subjects do not decide about the rule that classifies them as virtuous or free riders and restricts the possibility of punishment. As already emphasized, the specific legitimacy rule we selected, based on peer comparison, is very straightforward and formalizes a behavior that emerges spontaneously in the majority of cases. However, as a robustness check of the legitimacy principle, we also implement an alternative rule that is based on the average level of contribution in the group. Ertan et al. (2008) show that a legitimacy rule based on the average might emerge endogenously when subjects vote to allow for unrestricted or restricted punishment. In our treatment, also this rule is exogenous: now, it entitles subjects to punish peers only if their own contribution has been higher than the average contribution in the group, and confers immunity to subjects whose contribution has been above the average. However, the average level is endogenously determined within the specific group. As in the Partial R. treatment, the feedback is limited to the contribution levels of less virtuous peers (here, peers who contributed less than the average and that, consequently, can be punished) and to the average level of contributions in the group. We ran two sessions of 20 subjects each and a total of 10 groups and we found no significant differences in terms of contribution levels, punishment behavior and earnings between this treatment and the Partial R. In both treatments,

the feedback on more virtuous peers is missing and the average seems to work as an anchor that drives subjects' contributions down.

## 5. Discussion and conclusion

In the experimental and theoretical literature on cooperation and punishment, the behavioral propensity (i) to cooperate with others at a personal cost and (ii) to punish non-cooperators even when it is personally costly in the long run has been termed *strong reciprocity* (see Gintis, 2008). As Fudenberg and Pathak (2010) point out, understanding *when* and *why* costly punishment actually facilitates cooperation in public goods games is important both for the design of economic institutions and for modelling the evolution of cooperation. Our work contributes to shed light on the issue by means of a specially designed public goods game where punishment is allowed but only high contributors can punish and only low contributors can be punished. We wondered whether our legitimacy-based institution would be conducive to higher cooperation levels, compared to the *VCM* with unrestricted punishment opportunities, despite the lack of additional monetary incentives to cooperate.

Our results confirm that this is the case, providing clear evidence that legitimate punishment yields substantial benefits to cooperation<sup>18</sup>. Further, it leads to significantly higher earning levels, in the aggregate. Therefore, legitimate punishment turns out to be a more successful sanctioning institution along both dimensions of efficiency: contribution and earning levels<sup>19</sup>. We show that an exogenous institution can work extremely well by allowing for peer punishment to occur, but at the same time suitably restricting access to it.

We claim that an important implication to be drawn regards the classic principal-agent problem. In the standard analysis of the principal-agent relationship, principals hire agents due to the efficiency gains conferred by delegation. However, principal-agent relationships are typically characterized by a conflict of interest and asymmetric information. Falk and Kosfeld's (2006) well-known results indicate that there are "hidden costs" of control, as the decision to control significantly reduces the agents' willingness to act in the principal's interest: explicit incentives backfire and performance is lower if the principal controls, compared to if he trusts. As the two

---

<sup>18</sup> Also Ertan et al. (2008) find that an institution based on prohibiting punishment of high contributors is effective in raising cooperation levels and earnings. However, unlike the present study, they (1) focus on an average-based rule (the one we considered in the extension illustrated in section 4.4) and (2) investigate the dynamics of its endogenous emergence (through voting) when several institutional options are available *ex ante*.

<sup>19</sup> In light of these results, we view our findings as supportive of evolutionary models based on group selection such as Boyd et al. (2003), where the possibility that punishment not only fosters cooperation but also raises group average payoffs plays a critical role.

authors point out, “Elements in the labor contract that can be perceived as signals of distrust and control, such as minimum performance requirements, may harm more than help. Similarly, characteristics of the workplace environment that limit freedom of choice and signal low expectations, such as high levels of monitoring and surveillance, may be equally counterproductive” (p. 1612). Further, the free riding problem which characterizes public good provision and team working (Alchian and Demsetz, 1972) emerges. Therefore, it is natural to ask the following question: what monitoring and incentive schemes can be designed in order to enable the advantages of delegation to be realized? We argue that legitimate punishment provides a satisfactory answer, as it represents an enforcement device which is at the same time decentralized – because the enforcement of cooperation is delegated to the members of the group –, legitimate – because a member of a group can punish another member only if her contribution is higher than the contribution of the member she wants to punish –, and efficient – because it leads to higher levels of cooperation and earnings.

On the whole, our three-treatment design reveals that it is the *interaction* between behavioral restrictions and amount of information that crucially affects aggregate cooperation levels and earnings. The significant difference between contribution levels in the Full R. and the Partial R. treatment (Result 2) indicates that providing subjects with explicit information about higher contributors’ choices, that is ‘virtuous’ subjects’ behavior, plays a key role in the enforcement of cooperation. In this regard, it is natural to refer to an interesting series of recent experimental articles investigating the role of ‘leadership’ in social dilemma games and finding that leadership significantly raises average contribution levels. In these studies, leadership is typically implemented by letting an appointed leader influence others ‘by example’: she decides and announces her contribution *before* the other group members (simultaneously) make their contribution decisions (Van der Heijden and Moxnes, 2003; Gächter and Renner, 2004; Güth et al., 2007). In contrast, in our work we impose all subjects’ contribution and sanctioning decisions to occur simultaneously in every period. Further, higher contributors’ choices are never made salient throughout the experiment. However, the significant difference in contribution levels observed between our Full R. and Partial R. treatment suggests the following interpretation: subjects behave as if they perceived the legitimacy principle as endogenously conferring a leadership to high contributors, by making them (and only them) entitled to sanction lower contributors and (at least partially) immune from sanctioning. Under full information, this form of endogenous leadership (through legitimacy) leads to a significant increase in average contribution

levels, in line with the aforementioned leadership papers<sup>20</sup>. An even more specific analogy connects our paper to Güth et al.'s (2007) experiment, where, in one of the implemented treatments, they suppose that full information holds and leaders can punish others through exclusion, i.e. veto power. Interestingly, it is right in this case of an 'empowered leader' – the closest to our Full R. treatment – that they obtain the strongest result in terms of contribution levels, also compared to cooperation rates observed under pure leadership by example<sup>21</sup>.

Antisocial punishment is documented to play a relevant role when available: if the punishment activity is unrestricted, a non-negligible percentage of points are assigned also to subjects who contribute more than the punisher (weak antisocial punishment) and in some cases also to the most virtuous subjects (strong antisocial punishment). Under unrestricted sanctioning, the possibility that antisocial punishment occurs may also generate a 'motivational crowding-out' effect on virtuous subjects, as knowing that a significant probability to be punished exists even for high contributors may weaken their willingness to cooperate. By contrast, insofar as punishment is legitimate, this effect can be ruled out. More generally, a critical condition for a punishment institution to be successful is that "the incentives provided by punishment do not crowd out pre-existing social preferences that might have induced contributions in the absence of punishment, as is observed in a large number of public goods and principal agent experiments surveyed in Bowles (2008) and Bowles and Hwang (2008). The counterproductive effects of explicit incentives in the experiments they survey appear to arise when the punishment or fines fail to evoke shame in the shirker, but rather convey negative information about the individual imposing the incentive" (Carpenter et al., 2009). Our experimental result regarding the effectiveness of legitimate punishment suggests that, unlike under unrestricted punishment, the incentives provided by an institution based on legitimate punishment do not appear to crowd out pre-existing social preferences.

In general, the increase in contribution levels is stronger the higher the distance from the average. Information about the highest contributors affects the change in the levels of contribution of the most virtuous subject in the full information treatment with restrictions. In the full information treatment without restrictions, the information about the highest contributors is (opportunistically) interpreted by the less virtuous subjects as the assurance that someone else is

---

<sup>20</sup> As to empirical work, the effects of 'leading by example' have been analyzed with regard to charitable fundraising: a well-known result from these studies is that if renowned philanthropists donate to a specific project and this is publicly announced, others often tend to follow (Vesterlund, 2003). Further, so called 'seed money' typically generates a similar effect. We find that in our legitimacy-based framework a somewhat similar effect holds even though we refer to a simultaneous-move, rather than a sequential, game.

<sup>21</sup> As far as psychological experiments on leadership are concerned, it is interesting to note that several studies converge in finding a positive effect on contributions when the leader adheres to the principles of procedural fairness (see e.g. De Cremer et al., 2005).

carrying the burden of the public project, so that there is no need to do the same. Furthermore, as punishment is frequently used ‘unfairly’, all types of subjects in the baseline do not react to sanctions by enhancing their cooperative behavior in the next period.

The experiment run by Ertan et al. (2009) shows that people are willing to vote for an institution based on legitimate punishment. Our work can be seen as complementary to theirs, as the central question of our paper can be also expressed as follows: once subjects agree on a given legitimacy-based punishment institution, for the voluntary provision of a public good, does this institution work, with regard to both the achievement and sustainability of high cooperation and earning levels? How does it fare compared to its ‘natural benchmark’, that is an institution based on unrestricted punishment? Our experiment provides evidence that legitimate punishment can be an effective institution in deterring misconduct. Legitimate punishment is an ubiquitous phenomenon in several domains of real life, from access to positions in courts and police to family and intraorganizational relationships, education and political realms. Yet, so far there was no experimental evidence concerning the effects of exogenously determined legitimacy-based sanctioning institutions and feedback on cooperation.

Our study also leaves interesting avenues for further research, including the relative effectiveness of other legitimacy-based enforcement devices (e.g. based on positive incentives to cooperate, such as legitimate rewarding), the robustness of our major findings across alternative designs (e.g. ultimatum games, allowing for rejection only to responders who receive unfair offers) as well as the performance of the investigated mechanism across different cultural contexts. In this regard, we speculatively argue that legitimate punishment institutions might turn out to be even more effective within less developed societies than within industrialized ones, as recent research on cross-cultural differences (Herrmann et al., 2008; Gächter and Herrmann, 2010) indicates that the level of antisocial punishment here is far higher than in Western societies.

ACKNOWLEDGEMENTS: We gratefully acknowledge the Italian Ministry of Education, University and Research (2008 Prin project on “Social Capital, Corporate Social Responsibility and Performance” – Padova unit) and the Departments of Economics of the Universities of Trento and Verona for financial support. Our gratitude also goes to Angelo Antoci, Nicolò Bellanca, Alessandro Buccioli, Luigino Bruni, Domenico Colucci, Manolo Ferrante, Roberto Galbiati, Marcello Galeotti, Werner Güth, George Loewenstein, Luigi Luini, Nikos Nikiforakis, Stefania Ottone, Louis Putterman, Tiziano Razzolini, Laura Sabani, Mauro Sodini, Robert Sudgen, Erte Xiao and to the participants in the ESA World Meeting in Copenhagen, the VII LabSI Workshop on Experimental and Behavioral Economics in Siena, the 2nd Maastricht Behavioral and Experimental Economics Symposium, the IAREP/SABE 2010 Conference in Koln, the European Association of Law and Economics (EALE) Conference in Paris, and the colleagues of the Department of Economics of the Universities of Verona, Florence and Siena for useful suggestions and comments. The usual caveats apply.

## References

- Alchian, A. A., Demsetz, H. (1972). Production, Information Costs, and Economic Organization, *American Economic Review*, 62 (5), 777–795.
- Anderson, C., 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, 1-24.
- Andreoni, J. (1993). An experimental test of the public-goods crowding-out hypothesis, *American Economic Review*, 83, 1317-1327.
- Becker, G. (1968). Crime and punishment, *Journal of Political Economy*, 76 (2), 169-217.
- Botelho, A., Harrison, G.W., Pinto, L., Rutstrom, E.E. (2005). Social norms and social choice, Working Paper 30, Núcleo de Investigação em Microeconomia Aplicada (NIMA), Universidade do Minho.
- Boyd, R., Gintis, H., Bowles, S., Richerson, P.J. (2003). The evolution of altruistic punishment, *Proceedings of the National Academy of Science of the United States of America*, 100, 3532-3535.
- Carpenter, J., Bowles, S., Gintis, H., Hwang, S. (2009). Strong reciprocity and team production, *Journal of Economic Behavior and Organization*, 71 (2), 221-232.
- Casari, M., Luini, L. (2009). Cooperation under alternative punishment institutions: an experiment, *Journal of Economic Behavior and Organization*, 71(2), 273-282.
- Casari, M., Plott, C. (2003). Decentralized management of common property resources: experiments with a centuries-old institution, *Journal of Economic Behavior and Organization*, 51, 217-247.

Chan, K.S., Godby, R., Mestelman, S., Muller, R.A. (2002). Crowding-out voluntary contributions to public goods, *Journal of Economic Behavior and Organization*, 48, 305-317.

Chen, Y., Plott, C.R. (1996). The Groves-Ledyard mechanism: An experimental study of institutional design, *Journal of Public Economics*, 59, 335-364.

Cinyabuguma, M., Page, T., Putterman, L. (2006). Can second-order punishment deter perverse punishment?, *Experimental Economics*, 9 (3), 265-279.

De Cremer, D., van Knippenberg, D., van Knippenberg, B., Mullender, D., Stinglhamber, F. (2005). Rewarding leadership and fair procedures as determinant of self-esteem, *Journal of Applied Psychology*, 90, 1, 3-12.

Denant-Boemont, L., Masclet, D., Noussair, C.N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment, *Economic Theory*, 33, 145-167.

Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A. (2008). Winners don't punish, *Nature*, 452, 348-351.

Egas, M., Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation, *Proceedings of the Royal Society: Biological Sciences*, 275, 1637.

Ertan, A., Page, T., and Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53, 495-511.

Falk, A., Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, 96, 5, 1611-1630.

Falkinger, J., Fehr, E., Gächter, S., Winter-Ebmer, R. (2000). A simple mechanism for the efficient provision of public goods: Experimental evidence, *American Economic Review*, 90, 247-264.

Fehr, E., Gächter, S. (2000). Cooperation and punishment in public goods experiments, *American Economic Review*, 90 (4), 980-994.

- Fehr, E., Gächter, S. (2002). Altruistic punishment in humans, *Nature*, 415, 137-140.
- Fehr, E., Rockenbach, B. (2003). Detrimental Effects of Sanctions on Human Altruism, *Nature*, 422, 137-140.
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments, *Experimental Economics*, 10, 171-178.
- Fudenberg, D., Pathak, P.A. (2010). Unobserved punishment supports cooperation, *Journal of Public Economics*, 94, 1-2, 78-86.
- Gächter, S., Herrmann, B. (2010). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia, *European Economic Review*, forthcoming.
- Gächter, S., Renner, E. (2004). Leading by example in the presence of free rider incentives, Working Paper, University of St. Gallen.
- Gächter, S., Renner, E., Sefton, M. (2008). The long-run benefits of punishment, *Science*, 322, 5907, 1510.
- Goette, L., Huffman, D., Meier, S., Sutter, M. (2010). Group membership, competition, and altruistic versus antisocial punishment: evidence from randomly assigned army groups, IZA Discussion Paper N. 5189.
- Gintis, H. (2008). Punishment and cooperation, *Science*, 319, 1345-1346.
- Gürerk, O., Irlenbusch, B., Rockenbach, B. (2006). The competitive advantage of sanctioning institutions, *Science*, 312, 108-111.
- Herrmann, B., Thoeni, C., Gächter, S. (2008). Antisocial punishment across societies, *Science*, 319, 1362-1367.
- Güth, W., Levati, M.V., Sutter, S., van der Heijden, E. (2007). Leading by example with and without exclusion power, *Journal of Public Economics*, 91, 1023-1042.

Kosfeld, M., Okada, A., Riedl, A. (2009). Institution formation in public goods games, *American Economic Review*, forthcoming.

Kurzban, R., Houser, D. (2005). Experiments investigating cooperative types in humans: a complement to evolutionary theory and simulations, *Proceedings of the National Academy of Sciences of the United States of America*, 102 (5), 1803-1807.

Ledyard, J. (1995). Public goods: a survey of experimental research, in Kagel, J., Roth, A. (eds.), *Handbook of Experimental Economics*, Princeton, Princeton University Press.

Masclot, D., Noussair, C., Tucker, S., Villeval, M.C. (2003). Monetary and non-monetary punishment in the voluntary contributions mechanism, *American Economic Review*, 93 (1), 366-380.

Nikiforakis, N., Engelmann, D. (forthcoming). Altruistic punishment and the threat of feuds, *Journal of Economic Behavior and Organization*,

Nikiforakis, N., Normann, H.T. (2008). A comparative statics analysis of punishment in public good experiments, *Experimental Economics*, 11, 4, 358-369.

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

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public goods experiments, *Games and Economic Behavior*, 68, 689 -702.

Noussair, C., Tan, F. (2009). Voting on Punishment Systems within a Heterogeneous Group, *CentER Discussion Paper N. 19*, Tilburg University.

Ones, U., Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation, *Journal of Economic Behavior and Organization*, 62(2), 465-521.

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

Page, T., Putterman, L., Unel, B. (2005). Voluntary association in public goods experiments: reciprocity, mimicry, and efficiency, *Economic Journal*, 115, 1032-1053.

Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D., Nowak, M.A. (2009). Positive interactions promote public cooperation, *Science*, 325, 1272-1275.

Rockenback, B., Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment, *Nature*, 444, 718-723.

Sethi, R., Somanathan, E. (1996). The Evolution of Social Norms in Common Property Resource Use, *American Economic Review*, 86 (4), 766-88.

Sutter, M., Haigner, S., Kocher, M. (2010). The carrot or the stick? Endogenous institutional choice in social dilemma situations, *Review of Economic Studies*, forthcoming.

van der Heijden, E., Moxnes, E. (2003). Leading by example? Investment decisions in a mixed sequential-simultaneous public bad experiment, *CentER Discussion Paper n. 38*, Tilburg University.

Vesterlund, L. (2003). Informational value of sequential fundraising, *Journal of Public Economics*, 87, 627-657.

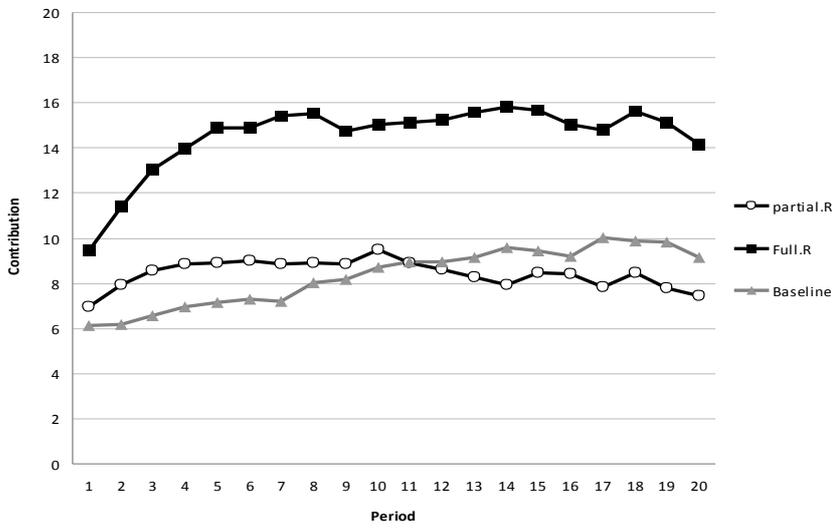
Xiao, E., Houser, D. (2010). Punish in public, *Journal of Public Economics*, forthcoming.

Xiao, E., Kunreuther, H. (2010). Punishment and cooperation in stochastic social dilemmas, mimeo.

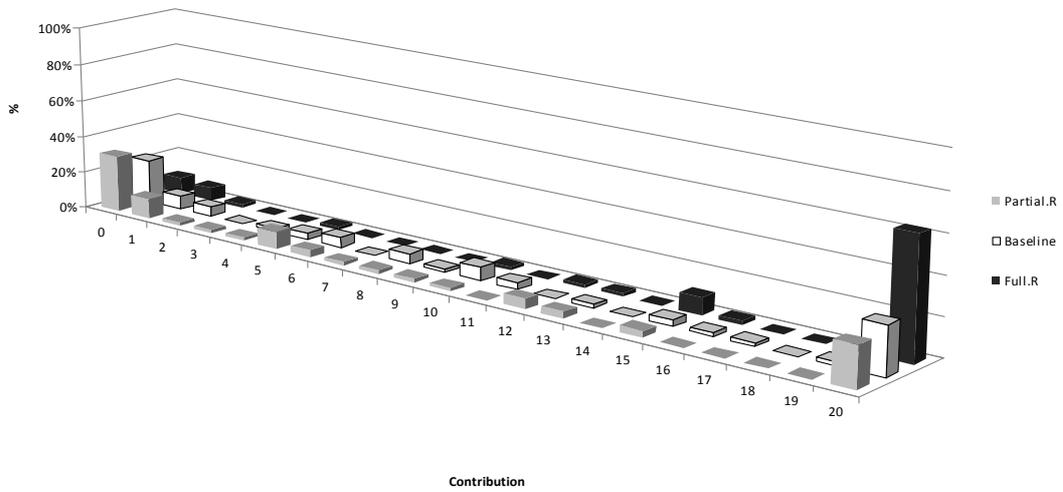
Yamagishi, T. (1986). The provision of a sanctioning system as a public good, *Journal of Personality and Social Psychology*, 51, 110-116.

# FIGURES AND TABLES

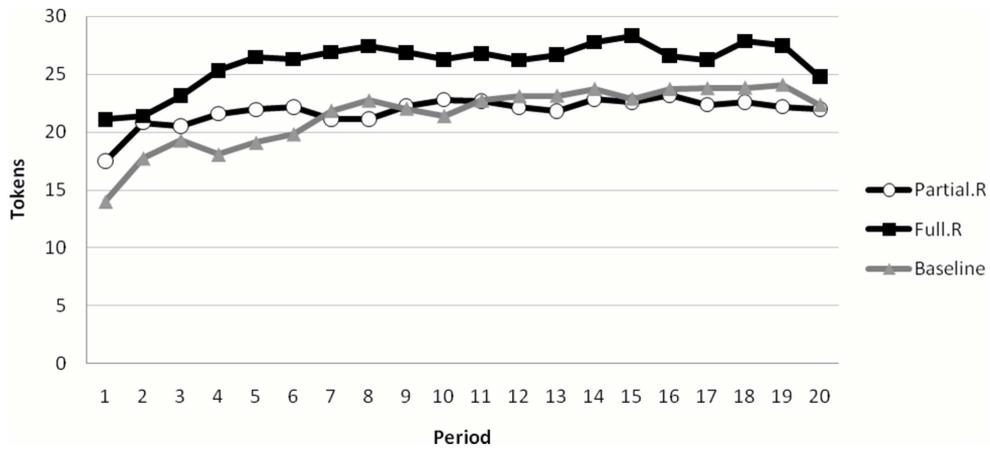
*Figure 1. Average contributions*



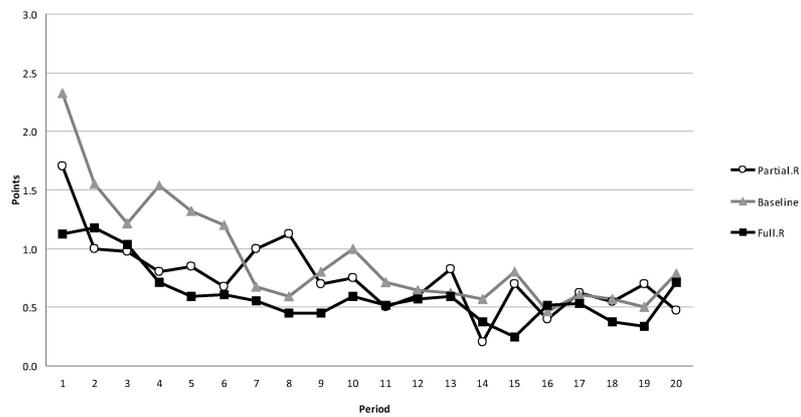
*Figure 2. Distribution of contributions in the final period*



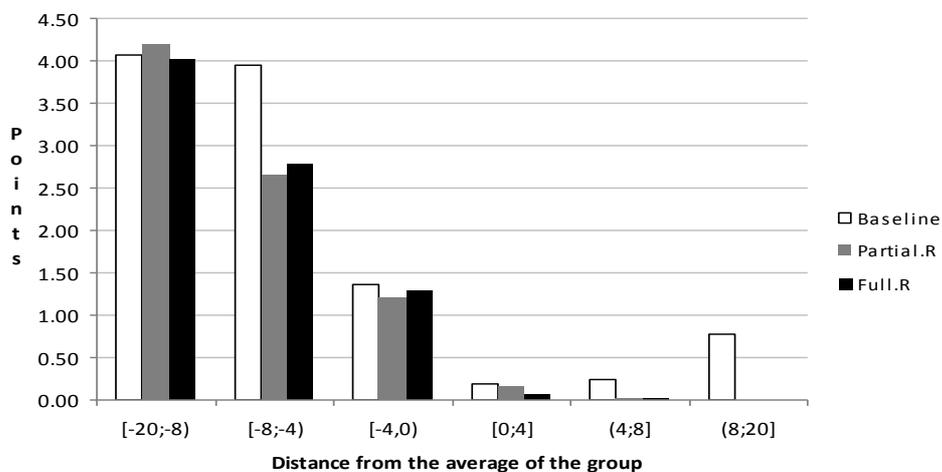
**Figure 3. Average earnings (tokens)**



**Figure 4. Average quantity of points given**



**Figure 5. Average quantity of points received as a function of the distance from the average of the group**



**Table 1. Treatments**

		Restrictions	
		Yes	No
Information	Full	Full.R (3 sessions; 14 groups; 56 subjects)	Baseline (3 sessions; 14 groups; 56 subjects)
	Partial	Partial.R (3 sessions; 14 groups; 56 subjects)	

**Table 2. Cost function**

<b>Points</b>	0	1	2	3	4	5	6	7	8	9	10
<b>Cost</b>	0	1	2	4	6	9	12	16	20	25	30

Table 3. Mean contribution

<b>Group</b>	<b>Baseline</b>	<b>Partial.R</b>	<b>Full.R</b>
1	13.76 (7.34)	16.06 (3.26)	18.69 (3.40)
2	18.40 (3.31)	4.05 (3.75)	10.76 (2.19)
3	4.94 (1.32)	3.43 (1.59)	16.90 (2.38)
4	11.30 (4.03)	5.44 (0.85)	0.81 (0.33)
5	12.85 (4.04)	1.74 (2.21)	18.34 (1.67)
6	4.58 (2.95)	9.59 (1.85)	17.70 (3.51)
7	6.46 (0.85)	19.20 (2.20)	5.69 (3.54)
8	2.18 (0.72)	3.70 (2.79)	18.91 (2.68)
9	4.39 (2.37)	14.50 (2.03)	16.85 (3.42)
10	1.64 (0.94)	3.10 (1.01)	18.95 (2.28)
11	2.84 (2.11)	15.06 (5.37)	13.73 (4.63)
12	15.13 (4.77)	9.53 (1.97)	14.38 (3.24)
13	7.05 (1.63)	8.63 (3.39)	13.56 (1.96)
14	11.11 (2.95)	3.89 (1.17)	18.09 (2.29)
Mean	8.42	8.33	14.53

Standard deviations in parentheses

Table 4. Determinants of contribution

Contribution	Random Effect Tobit
Partial.R	-5.18 *** (1.60)
Baseline	-4.94*** (1.70)
Av_First	1.17*** (0.24)
Constant	-6.00 (6.22)
Log-likelihood	-7677.25
Chi(2)	95.87
N. of obs.	3360

The dependent variable takes values from 0 to 20. *Av\_first*: group average contribution in the first period. *Baseline*: dummy variable taking value 1 if the treatment is the baseline treatment; *Partial.R* dummy variable taking value 1 if the treatment is the baseline Partial.R treatment; Controls: age, nationality, major, gender and number of experiments in which the subject has been involved in the past.

\*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%.

Table 5. Average point given per period

<b>Group</b>	<b>Baseline</b>	<b>Partial. R.</b>	<b>Full.R</b>
1	0.39 (0.59)	0.93 (1.05)	0.31 (0.83)
2	0.53 (1.04)	0.78 (0.52)	0.93 (0.78)
3	0.03 (0.08)	0.25 (0.43)	0.48 (0.57)
4	0.56 (0.53)	0.54 (0.53)	0.26 (0.27)
5	0.75 (0.89)	0.85 (0.82)	0.58 (0.73)
6	0.90 (1.05)	1.06 (1.33)	0.79 (2.10)
7	0.31 (0.62)	0.18 (0.46)	0.48 (1.29)
8	0.56 (0.45)	0.98 (1.06)	0.31 (0.76)
9	1.66 (1.10)	1.20 (0.85)	0.60 (0.67)
10	2.26 (0.81)	0.83 (0.49)	0.55 (0.93)
11	1.65 (1.04)	0.53 (0.51)	0.60 (0.39)
12	0.19 (0.25)	0.69 (0.63)	1.18 (0.75)
13	1.63 (0.77)	1.53 (0.78)	1.06 (0.77)
14	1.54 (0.37)	0.53 (0.47)	0.34 (0.36)
Mean	0.77	0.93	0.60

Standard deviations in parentheses

Table 6. Antisocial punishment (A.P.) in the Baseline

	(1) Points given  by <i>i</i> to <i>j</i>	(2) Weak A.P.  $\text{Contr}_i < \text{Contr}_j$	(3)  % Weak a.p.	(4) Strong A.P.  $\text{Contr}_j > \text{Contr}_i$ and $\text{Contr}_j > \text{AV\_contr}$	(5)  % Strong A.P.
<b>Group</b>					
1	29	5	17.2%	5	17.2%
2	42	0	0.0%	0	0.0%
3	2	0	0.0%	0	0.0%
4	45	12	26.7%	8	17.8%
5	59	10	16.9%	3	5.1%
6	72	29	40.3%	16	22.2%
7	25	0	0.0%	0	0.0%
8	45	2	4.4%	1	2.2%
9	133	5	3.8%	3	2.3%
10	181	48	26.5%	36	19.9%
11	132	18	13.6%	7	5.3%
12	15	0	0.0%	0	0.0%
13	130	66	50.8%	41	31.5%
14	123	7	5.7%	6	4.9%
<b>Total</b>	<b>1033</b>	<b>202 (19.5%)</b>		<b>126 (12.2 %)</b>	
<b>Mean</b>	<b>73.79</b>	<b>14.43</b>	<b>14.71%</b>	<b>9.00</b>	<b>9.17%</b>

Table 7: Determinants of the quantity of punishment points received

<b>Received points</b>	<b>Baseline</b>	<b>Partial.R</b>	<b>Full.R</b>
Positive distance from average	0.04 (0.047)	-0.78 *** (0.11)	-0.97*** (0.16)
Absolute negative distance from average	0.91*** (0.047)	0.61 *** (0.041)	0.64*** (0.03)
Constant	-0.60 (2.61)	-3.43* (1.83*)	-1.25 (1.93)
Log-likelihood	-1154.05	-1128.12	-874.83
Wald Chi(2)	397.69	332.10	389.15
N. Of obs.	1120	1120	1120

Random Effect Tobit.

The dependent variable takes values from 0 to 30.

Positive distance from average is the difference between subject's contribution and the average contribution of the group; it takes value equal to zero when the subject contributes less than the average. Absolute negative distance from average is the difference between average contribution of the group and subject's contribution; it takes value equal to zero when the subject contributes more than the average. Controls: age, nationality, major, gender and number of experiments in which the subject has been involved in the past.

Standard errors in parentheses

\*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%.

Table 8. Determinants of the change in contribution levels

Contribution at t - contribution at t-1	Below the average in t-1			Not below the average in t-1		
	Baseline	Partial.R	Full.R	Baseline	Partial.R	Full.R
Distance from average at t-1	-0.60*** (0.08)	-0.50*** (0.07)	-0.78*** (0.08)	-0.78*** (0.06)	-0.64*** (0.06)	-0.42*** (0.06)
Points received at t-1	0.10 (0.08)	0.50*** (0.09)	0.60*** (0.15)	-0.86*** (0.22)	-0.02 (0.30)	0.85** (0.44)
Constant	1.70 (2.77)	-0.11 (3.33)	-3.81 (3.00)	-1.46 (2.15)	3.13 (2.33)	-0.90 (1.92)
Log-likelihood	-1165.19	-1194.05	-931.65	-1490.15	-1550.95	-1929.30
Wald Chi(2)	82.57	151.54	190.99	201.26	103.62	52.85
N. Of obs.	468	456	329	596	608	735

The dependent variable takes values from -20 to 20.

Distance from average at t-1 is the difference between subject's contribution at t-1 and the average contribution of the group at t-1.

Controls: gender, age, nationality, Controls: age, nationality, major, gender and number of experiments in which the subject has been involved in the past.

Standard errors in parentheses

\*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%.

Table 9. Impact of information about highest contributions in the group.

<b>Contribution at t - contribution at t-1</b>	<b>Below the average in t-1</b>		<b>Not below the average in t-1</b>	
	<b>Baseline</b>	<b>Full.R</b>	<b>Baseline</b>	<b>Full.R</b>
Distance from average at t-1	-0.83*** (0.15)	-0.66*** (0.16)	-0.52*** (0.19)	-1.63*** (0.17)
Distance from the highest contribution at t-1	-0.12** (0.064)	0.08 (0.10)	-0.06 (0.10)	0.26** (0.12)
Points received at t-1	0.08 (0.08)	0.59*** (0.15)	-0.51 (0.33)	0.09 (0.42)
Constant	1.77 (2.45)	-3.63 (2.98)	5.64 (3.27)	8.81 (3.03)
Log-likelihood	-1163.29	-931.33	-367.37	-306.41
Wald Chi(2)	87.93	191.74	22.18	109.38
N. Of obs.	468	329	152	127

The estimation is limited to the sub-sample of subjects whose contribution in the previous period was not the highest of the group. The dependent variable takes values from -20 to 20.  
 Distance from average at t-1 is the difference between subject's contribution at t-1 and the average contribution of the group at t-1.  
 Distance from highest at t-1 is the absolute difference between subject's contribution at t-1 and the highest contribution of the group at t-1.  
 Controls: age, nationality, major, gender and number of experiments in which the subject has been involved in the past.

\*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%.