Legitimate Punishment, Immunity, and the Enforcement of Cooperation

Marco Faillo [a] Daniela Grieco[b] Luca Zarri[c]

(Very preliminary and incomplete version. Please do not quote or circulate)

Sep 20, 2010

Abstract

In the framework of a finitely repeated public goods game with costly punishment options, we introduce a novel restrictive setup where a principle of legitimacy *holds, in the sense that only virtuous behavior (that is, being a high contributor) allows one to gain access to sanctioning opportunities ('entitlement') and only wrongful behavior (that is, being a low contributor) makes one a potential target of peer punishment ('desert'). As a consequence, acting virtuously guarantees that it will not be possible to be punished by less virtuous subjects ('immunity'). These restrictions, by allowing for 'legitimate punishment' only, rule out by construction so called antisocial punishment as well as vengeful behavior. Moreover, we manipulate the amount of information over others' contributions that subjects receive before making their punishment decisions. Our preliminary results show that restrictions* per se *do not affect the cooperation levels; by contrast, virtuous restrictions combined with feedback on virtuous peers significantly increase contribution levels and make cooperation sustainable over time.*

JEL Classification: C72; C91; C92; D23 ; D72.

Keywords: Experimental Economics; Public Good Games; Costly Punishment; Cooperation; Legitimacy; Immunity.

[a] marco.faillo@unitn.it, Department of Economics, University of Trento.

[b] daniela.grieco@univr.it, Department of Economics, University of Verona.

[c] 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. In schools, only teachers are allowed to sanction bad students. At home, it is parents only who can do the same with their children. 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 drugs or prostitutes, 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 such a principle 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[1].

[1] On philosophical grounds it can be plausibly maintained that the very existence of the modern state itself

In this paper, we investigate – to our knowledge for the first time – this legitimacy-punishment nexus experimentally within a public goods game framework, in order to see whether legitimate punishment turns out to be an effective institution in enforcing cooperation. In doing this, we depart from most of the existing experimental literature on punishment, as, from the seminal works of Ostrom et al. (1992) and Fehr and Gaechter (2000; 2002) onwards, 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 Gaechter'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 shows that there is also a 'dark side' of unrestricted punishment. In particular, the following drawbacks have been identified in the last years. First, since everyone is free to punish everyone else, 'antisocial' punishment – that is, low contributors punishing high contributors – often arises and reduces contribution rates, especially within less advanced societies (Herrmann et al., 2008). Second, when multiple stages of punishment are allowed, counterpunishment, second-order perverse punishment and feuds are likely to be triggered, leading to a demise of cooperation (Cinyabuguma et al., 2006; Denant-Boemont et al., 2007; Nikiforakis, 2008 and Nikiforakis and Engelmann, 2008). This shows that unrestricted punishment is not robust, as an effective cooperation enforcement device, to institutional changes. Third, a further problem with this form of punishment is that it exclusively relies on deterrence, that is on *extrinsic* motives to cooperate. Since it turns out that people are often intrinsically motivated to comply, the risk is either not to

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.

elicit such intrinsic motivations or to even crowd them out, especially when incentives are weak[2]. The key point here is that, as prominent scholars in law and economics have pointed out, one important function of sanctioning is to *express norms* (Cooter, 1998; Sunstein, 1996; Tyran and Feld, 2006). In this regard, our work, by focusing on 'virtuously restricted', legitimate punishment, is close in spirit to experimental papers relying on the expressive power of norms, obligations (Galbiati and Vertova, 2008) and moral appeals (Dal Bo' and Dal Bo', 2009). Xiao and Houser (2010) confirm that norm salience has an effect on cooperation behavior, as their data support the hypothesis that public punishment promotes norm-obedience by reinforcing cooperation norms to both the punishee and to those who observe punishment. Dal Bo' and Dal Bo' (2009) find that, when combined with punishment options, 'moral appeals' – that is, observing a message with a moral standard – significantly increase contribution rates[3]. As argued by Bohnet and Cooter, a law may elicit intrinsic motivation by framing an act as wrong. Analogously, we explore the 'institutional moral framing' effect of a punishment institution where acting wrongfully does not give access to punishment options and makes one vulnerable to virtuous subjects' punishments.

2. Related literature

In a public goods game or voluntary contribution mechanism (*VCM*) framework, there is a group of subjects who, as the game starts, receive an individual monetary

[2] An example of this phenomenon is provided by the well-known experiment run by Gneezy and Rustichini (2000), showing that when fined for arriving late to their child's day care, the number of late-coming parents increased.

[3] In a similar vein, Croson and Marks (2001) investigate the impact of *recommendations* (without appealing to moral rules) on contributions in public goods games, but find limited effects. Dale and Morgan (2004) found that recommendations favoring the top contribution were less effective than recommendations favoring intermediate contributions.

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, standard prediction based on the canonical model of *Homo Oeconomicus* is that everyone should free ride, both in the one-shot and in the repeated game. However, in the finitely repeated version, 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. Gurerk et al., 2006 and Kosfeld et al., 2009), 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; Fudenberg and Pathak, 2009; Rand et al., 2009). In their well-known and often cited 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, by having access to so called 'costly' or 'altruistic' punishment, 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[4]. The emergence of voluntary, costly punishment in the laboratory is a puzzle for

[4] Analogously, the introduction of explicit punishment and/or rewarding opportunities significantly affects subjects' behavioral choices in the experimental games studied by Fehr et al. (1997) and Fehr and Rockenbach (2003).

standard economic theory, as it is in contrast with the idea that subjects act selfishly in order to systematically maximize their monetary gains. In other words, in a finitely repeated *VCM* with punishment options, the 'canonical model' recalled above predicts that subjects *will not* use such options, due to the net monetary costs associated with their usage[5]. By contrast, despite the seemingly irrational nature of this behavioral attitude, 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; Nikiforakis and Normann, 2008; Anderson and Putterman, 2006; Rockenback and Milinski, 2006) and with fixed groups playing a finite number of times (Yamagishi, 1986; Fehr and Gächter, 2000; Page et al., 2005). Experimentally, it has been shown to represent a powerful decentralized enforcement device, through which it is possible to induce and successfully sustain cooperation in social dilemmas.

Like these studies, in this paper we focus on a decentralized mechanism based on exogenously imposed sanctioning opportunities. However, unlike existing work, we suitably restrict access to punishment options: our sanctioning institution is based on 'legitimate punishment' in the sense that it prescribes that only 'high contributors' can sanction and only 'low contributors' can be sanctioned. As we pointed out in the introduction, recent lab studies on unrestricted punishment show that undesirable behaviors – such as (first-order) antisocial punishment and, when multiple stages of punishment are allowed, feuds, counterpunishment and higher-order perverse punishment – often occur and, over time, lead to a demise of cooperation. In contrast, an institution based on legitimate punishment rules out these forms of detrimental punishment by construction. On positive grounds, it may also elicit people's intrinsic motivation to contribute and punish low contributors. Hence, legitimate punishment, compared to

[5] 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.

unrestricted punishment, may act as a sanctioning mechanism which has expressive, norm-activating effects (like a 'mild sanction', in the law and economics terminology). Does a legitimate punishment institution have norm-activating effects, in a public goods game framework? Answering this question is among the goals of this paper.

3. Experimental setup

In our sanctioning institution, some key restrictions are exogenously imposed with regard to both *who* is allowed to punish and *whom* can be punished[6]. 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 (principle of *legitimacy*).

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

[6] Therefore, our design also differs from recent experimental *VCM* protocols where norms prescribing who can punish and/or whom 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.

be a punisher at stage 2[7]. More specifically, we assume that a subject is entitled to punish another subject at stage 2 only if her contribution at stage 1 has been *higher* than the contribution of the peer she wants to punish. As a consequence, high contributors are (partially) immune from punishment, in the sense that they do not deserve to be sanctioned by players who contributed less then 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 92 subjects participated voluntarily in the experiment at the CEEL Lab of the University of Trento. A total of 5 sessions were conducted, between December 2009 and June 2010. Three sessions had 20 participants and the other two sessions had 16 participants. The experiment was programmed by using the z-tree platform (Fischbacher, 2007). The subjects were undergraduate students of the University of Trento. The subjects were undergraduate students (63% from Economics, 53% females, 80% 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 13.94 euros and the sessions averaged approximately 1 hour and 30 minutes. The instructions were read aloud by the

[7] 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 confers to them a right to costly punish wrongdoers. This is an important difference, as we are interested in analyzing the moral framing power of a punishment institution, rather than its deterrence power on the basis of monetary incentives only.

experimenter and 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 do this as "repeated interaction is a common feature of naturally occurring environments (e.g., businesses or collectives) in which punishment often 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[8]. The design and the parametric structure of the experiment are based on those of Fehr and Gächter (2000).

[8] 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 (2004), 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). In contrast, Fehr and Gaechter (2000) find differences in outcomes between partner and stranger protocol in their *VCM* experiment.

3.2. Treatments

We implemented three experimental treatments: a baseline, unrestricted punishment and full information (UF) treatment, a restricted punishment with full information (RF) treatment and a restricted punishment with partial information (RP) treatment.

There were 2 sessions (one with 20 subjects and the other with 16 subjects) for the UF, 1 session (with 16 subjects) for the RF and 2 sessions (with 20 subjects in each) for the RP. 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).

[Table 1 here]

Behavioral restrictions and feedback about others' contribution levels in the group. In the UF 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 Gaechter 2000), where everyone can freely punish everyone else in the group. We used it as our baseline. The other two treatments are both based on *legitimacy* (i.e. entitlement and desert): both in the RF and the RP treatment, a subject 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 RF subjects are informed about the full vector of others' contributions (like in UF), in RP 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. The reason why we differentiate the two legitimacy-based treatments in terms of feedback is that manipulating the feedback that players receive about others' behavior seems to affect their own behavior in *VCM* environments (see e.g. Nikiforakis, 2008; Xiao and Houser, 2010).

3.2.1. Legitimacy-based treatments

While our UF (baseline) treatment is based on the standard *VCM* with punishment options (Fehr and Gächter, 2000), our two legitimacy-based treatments (RF and RP) share the following features. In stage 1, at the beginning of each period each participant receives a fixed amount e =20 of tokens. 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)

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[9]. 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 here]

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

[9] It is important to make clear that we never used loaded terms such as 'legitimacy', 'entitlement', 'desert' and 'immunity' during the experiment (see the Instructions, that will be added in the Appendix).

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 here]

Average contribution increases over time in UF and RF treatments, while it remains constant in the RP treatment.

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

[Table 3 here]

Besides this well known general positive effect of punishment, our data show (Figure 1) that, given the same type of restrictions on the punishment activity, subjects who are informed on the contributions of all the other members of their group (RF treatment) contribute more than subjects who are informed only on the average contribution of their group and on less virtuous peers' contributions (RP treatment). At the same time, given the same level of information, the introduction of restrictions on the punishment activity does not affect the level of contribution.

This differences characterize also the final period of the game (Figure 2). Results 2 and 3 follow.

[Figure 2]

Result 2. The introduction of restrictions per se *does not increase the level of cooperation.*

Result 3. The provision of detailed information on the levels of contribution within the group results in a higher level of cooperation.

These results are supported by the regression analysis reported in Table 4. Testing the following equation sheds further light on the role of restrictions and information in shaping the contribution levels:

$$contribution_{igt} = \beta_0 + \beta_1\, av\_first_g + \beta_2\, treatment + controls_i + \alpha_t + \theta_g + \varepsilon_{igt} \quad (Eq.\ 1)$$

where av_first_g is the group average level of contribution in the first period and *treatment* is the dummy variable identifying the treatment (UF, RF, RP). The $controls_i$ involve socio-economic and demographic variables like gender, age, nationality; α_t and θ_g are the fixed effects by period and by group respectively.

[Table 4 here]

Table 4 shows that contributions in each period are positively (and significantly) affected by the average contribution in the group in the first period. Therefore, each group's behavior in the first period represents a key determinant of next contribution choices in the group: cooperation seems to be sustained by idiosyncratic features of the specific group that could not be based on any form of indirect feedback and reputation.

3.2. Punishment behavior

What it is particularly interesting about Result 2 is that introducing restrictions in the aim of preventing the assignment of punishment points to virtuous subjects does not result in higher levels of contribution. 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 UF treatment.

[Figure 3 here]

With regard to the distribution of punishment points, in all the treatments we observe the typical decreasing pattern, which is faster in the RF treatment (Figure 3). In the RF treatment subjects assign, on average, a smaller quantity of punishment points (see Table 5).

[Table 5 here]

However, the difference between the average quantity of points assigned in the three treatments is not statistically significant. The determinants of the amount of punishment points given by each subjects are investigated by testing the following equation:

$$punishment\ points\ given_{igt} = \beta_0 + \beta_1\ av\_first_g + \beta_2\ treatment + \beta_3\ dist\_av_{igt} + controls_i + \alpha_t + \theta_g + \varepsilon_{igt} \qquad (Eq.\ 2)$$

where variables are defined as in *Eq. 1* and $dist\_av_{igt}$ represents the distance between the punisher's contribution level and the average contribution level in the group.

Table 6 shows that treatments do not play any role in shaping the number of points assigned, whereas having exhibited a virtuous behavior with respect to the group (i.e. by how much a subject's contribution has been above the average) significantly affects punishment levels.

[Table 6 here]

In the UF treatment a non-negligible percentage of punishment points are assigned to virtuous subjects. Table 7 reports the absolute quantities (column 2) and the percentage (column 3) of punishment points assigned in the UF treatment by a subject *i* to a subjects *j* when the contribution of *i* is smaller than the contribution of *j*. We define 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 14.4% of the overall punishment activity (number of punishment points assigned in all periods) can be classified as weak-antisocial, while 8.9% is strongly antisocial.

[Table 7 here]
[Figure 4 here]

The presence of a mild form of punishment of virtuous subjects (strong antisocial punishment) in the UF treatment emerges also in Figure 4, which displays the relation between the distance from the average contribution of the group and the average quantity of points received. In the UF treatment, in some cases strong positive deviations are still punished. This evidence is supported by the results of the following regression analysis:

$$punishment\ points\ received_{igt} = \beta_0 + \beta_1\ pos\_dev\_\ av_{igt} + \beta_2\ neg\_dev\_\ av_{igt} + \beta_3\ controls_i + \alpha_t + \theta_g + \varepsilon_{igt} \qquad (Eq.\ 2)$$

whose results are reported in Table 8. Variables *pos_dev_ av*$_{igt}$ and *abs_neg_dev_ av*$_{igt}$ capture the distance between the subject's contribution and the average contribution in the group when this distance is positive and negative (in this last case, the difference is taken in absolute terms), respectively.

[Table 8 here]

While in all treatments the quantity of punishment points received decreases as the negative deviation from the average increases, positive deviation 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).

3.3. Determinants of changes in individual contribution levels

As sections above have shown that treatments UF and UR are significantly different in terms of contributions levels, but not in terms of punishment points assigned,

an analysis of the efficacy of punishment in modifying contribution levels is needed. The following equation is tested:

$$contribution_{igt} - contribution_{igt-1} = \beta_0 + \beta_1\, received\_punishment_{igt-1} + \beta_2\, dist\_av_{igt-1} + \beta_3\, treatment + controls_i + \alpha_t + \theta_g + \varepsilon_{igt} \quad (Eq.\ 3)$$

where $received\_punishment_{igt-1}$ represents the number of punishment points that the subjects has received in the previous period, whereas $dist\_av_{igt-1}$ is the distance between the subject's contribution and the average contribution in the group in the previous period.

Table 9 shows that both variables turn out to be significant in all the treatments: the more a subjects has been sanctioned in a period, the higher the increase in her contribution in the next period, no matter the presence of restrictions in punishment or the feedback she receives. Furthermore, the higher the distance to the average contribution in the group, the stronger this reaction. Figure 4 show a regression towards the mean.

[Table 9 here]

Result 5. Punishment is effective in all the treatments and the increase in contribution levels is stronger the higher the distance from the average and regresses to the mean.

In the aim of exploring the role of feedback in shaping contribution reaction to punishment, the equation tested above is modified by considering an additional

dependent variable, i.e. the distance between the subject's contribution and the highest contribution in the group (labeled $dist\_highest_{igt-1}$):

$$contribution_{igt} - contribution_{igt-1} = \beta_0 + \beta_1\, received\_punishment_{igt-1} + \beta_2\, dist\_av_{igt-1} + \beta_3\, dist\_highest_{igt-1} + \beta_4\, treatment + controls_i + \alpha_t + \theta_g + \varepsilon_{igt} \qquad (Eq.\ 4)$$

Table 10 documents the significant role of information on the most virtuous peer in affecting the increase in contribution in treatments UF and RF, i.e. in treatments where the full vector of peers' contribution is available and subjects have the possibility to use virtuous peers behavior as a reference point. In the RP treatment, subjects can confront with lower contributors only (although they receive the average contribution level in the group as well).

[Table 10 here]

Result 6. In full information treatments the highest contribution level in the group is used as reference point

4. Discussion and conclusion

In the literature, 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 called *strong reciprocity* (see Gintis, 2008). As Fudenberg and Pathak (2009) 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 modeling the evolution of cooperation. Our work contributes to shed light on the issue by means of a specially designed public goods game where *only high contributors can punish* and *only low contributors can be punished*. Since several *VCM* studies on punishment have shown that it is mainly high contributors who punish low contributors, we expected that our institution – by channelling subjects' sanctioning attitudes towards the attainment of a socially beneficial goal (i.e. sustainable high contribution levels) through 'virtuous restrictions' on both who can punish and whom can be punished – would be conducive to higher cooperation levels, compared to both the standard *VCM* (with no punishment opportunities) and the *VCM* with (unrestricted) punishment opportunities. We expected our mechanism to deter misconduct also in the light of the fact that such an institution centred around legitimacy and immunity – and, therefore, rendering strong reciprocity institutionally salient – completely rules out antisocial punishment and punishment motivated by vengeance (see Mocan, 2008), that is the forms of sanctioning that so far turned out to be mostly detrimental to the sustainability of cooperation (see Bochet et al., 2006).

Our experimental setup is explicitly built to test the role of restrictions in shaping contributions behaviour. Furthermore, the role of information on peers provided to (potential) punishers is investigated.

Our results show that punishment prevents the decline of cooperation over time in all the treatments and is effective in shaping contribution levels. Interestingly, it is the interaction between restrictions and information that affects cooperation levels.

When possible, antisocial punishment is documented to play a role: 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). However, restricting punishment opportunities is not enough to determine an increase in

cooperation, and we suspect that a crucial role is played by the amount of information legitimate punishers receive. In general, the increase in contribution levels is stronger the higher the distance from the average; in full information treatments the highest contribution level in the group is used as reference point. Peers' virtuous decisions seem to work as a "good example" and drive subjects' contribution level upward.

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 relationships, education and political realms. Yet there is no experimental evidence concerning the effects on legitimate punishment on cooperation. Our paper offers such evidence and investigates which conditions are critical to build virtuous institutions.

We expect legitimate punishment to work even better within less developed societies, as previous research on cross-cultural differences (Herrmann et al., 2008) has shown that the level of antisocial punishment here is far higher than in Western societies. Therefore, we predict that the higher the degree of antisocial punishment within a society, the better our mechanism works.

References

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.

Andreoni, J., Harbaugh, W., Vesterlund, L. (2003). The carrot or the stick: rewards, punishments, and cooperation, *American Economic Review*, 93 (3), 893-902.

Bochet, O., Page, T., Putterman, L. (2006). Communication and punishment in voluntary contribution experiments, *Journal of Economic Behavior and Organization*, 60, 11-26.

Bowles, S., Gintis, H. (2002). Behavioural science: Homo reciprocans, *Nature*, 415, 125-128.

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.

Burlando, R., Guala, F. (2005). Heterogeneous agents in public goods experiments, *Experimental Economics*, 8, 35-54.

Carpenter J., 2005, The demand for punishment, *Journal of Economic Behavior and Organization*.

Carpenter J., 2007, Punishing free riders: how group size affects mutual monitoring and the provision of public goods, *Games and Economic Behavior*, 60, 1, 31-51.

Carpenter J., Matthews P.H., 2007, What norms trigger punishment?

Casari, M., Luini, L. (2007). Peer punishment in teams: expressive or instrumental choice? *Working Paper*.

Casari, M., Luini, L., (). Cooperation under alternative punishment institutions: an experiment, *Journal of Economic Behavior and Organization*, forthcoming.

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). Crwoding-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.

Cooter, R.D., Expressive law and economics, *Journal of Legal Studies*, 27 (2), 585-608.

Croson, R., Marks, M. (2001). The effect of recommended contributions in the voluntary provision of public goods, *Economic Inquiry*, 39 (2), 238-249.

Dal Bo', E., Dal Bo', P. (2009). 'Do the right thing': the effects of moral suasion on cooperation, mimeo.

Dale, D.J., Morgan, J. (2004). Fairness equilibria and the private provision of public goods, mimeo UC Berkeley.

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.

Falkinger, J. (1996). Efficient private provision of public goods by rewarding deviations from average, *Journal of Public Economics*, 62, 413-422.

Falkinger, J. (2004). Noncooperative support of public norm enforcement in large societies, CESifo Working Paper N. 1368.

Falkinger, J., Fehr, E., Gächter, S., Winter-Ebmer, R. (2000). A simple mechanism fort he 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., Gächter, S., Kirchsteiger, G. (1997). Reciprocity as a Contract Enforcement Device: Experimental Evidence, *Econometrica*, 65 (4), 833-60.

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.

Fischbacher, U., Gaechter, S., Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment, *Economics Letters*, 71, 3, 397-404.

Fischbacher, U., Gaechter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments, *American Economic Review*, forthcoming.

Fudenberg, D., Pathak, P.A. (2009). Unobserved punishment supports cooperation, *Journal of Public Economics*, forthcoming.

Galbiati, R., Vertova, P. (2008). Obligations and cooperative behavior in public good games, *Games and Economic Behavior*.

Gintis, H. (2008). Punishment and cooperation, *Science*, 319, 1345-1346.

Gneezy, U., Rustichini, A. (2000a). Pay enough or don't pay at all, Quarterly Journal of Economics, 115 (3), 791-810.

Gneezy, U., Rustichini, A. (2000b). A fine is a price, Journal of Legal Studies, 29 (1), 1-17.

Gurerk, 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.

Isaac, R.M., Walker, J.M. (1988). Communication and free-riding behavior: the voluntary contributions mechanism, *Economic Inquiry*, 26, 585-608.

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.

Masclet, 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.

Mocan, N. (2008). Vengeance, NBER Working Paper N. 14131.

Nikiforakis, N., Engelmann, D. (2008). Feuds in the laboratory? A social dilemma experiment.

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

Nikiforakis, N. (2008). Punishment and counter-punishment in public geode games: Can we really govern ourselves?, *Journal of Public Economics*, 1(2), 91-112.

Oliver, P. (1980). Rewards and punishments as selective incentives for collective action: theoretical investigations, *American Journal of Sociology*, 85, 1356-1375.

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. (2000). Cheap talk and punishment in voluntary contribution experiments, Brown University Department of Economics Working Paper

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

Panchanathan, K., Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem, *Nature*, 432, 499-502.

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.

Sefton, M.R., Shupp, R., Walker, J. (2007). The effects of rewards and sanctions in provision of public goods, *Economic Inquiry*.

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

Sigmund, K., Hauert, C., Nowak, M.A. (2001). Reward and punishment, *Proceedings of the National Academy of Science of the United States of America*, 98, 10757-10762.

Smith, E.A., Bliege Bird, R. (2000). The benefits of costly signaling: Meriam turtle hunters, *Behavioral Ecology*, 14 (1), 116-126.

Sunstein, C.R., 1996, On the expressive function of law, University of Pennsylvania Law Review, 144, 2021-2031.

Tyran, J., Feld, L.P. (2006). Achieving compliance when legal sanctions are non-deterrent, *Scandinavian Journal of Economics*, 108 (1), 135-156.

Wiessner, P. (2003). The pleasures and perils of reward and punishment: strong reciprocity among the Ju/'hoansi Bushmen. Department of Anthropology, University of Utah.

Xiao, E., Houser, D. (2010). Punish in Public, 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 contribution

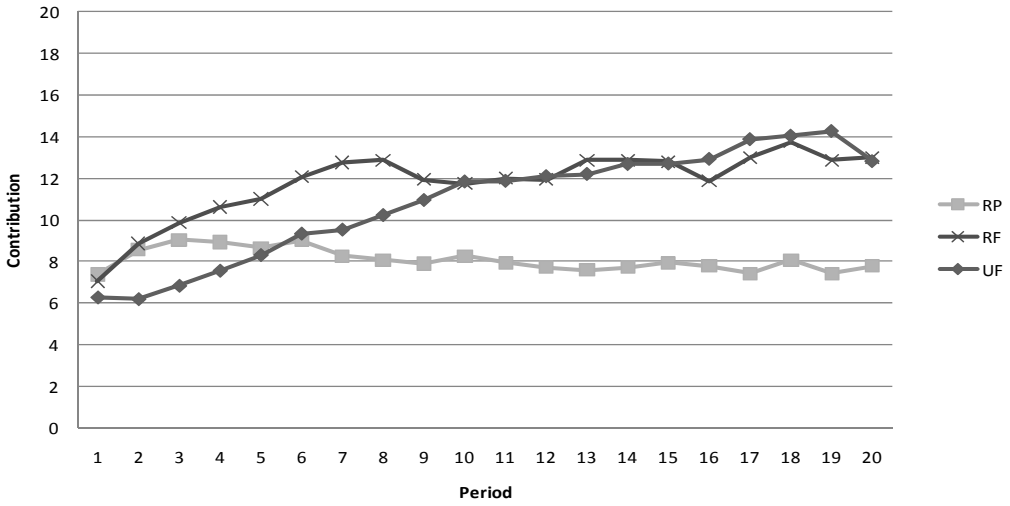

Figure 2. Distribution of contributions in the final period.

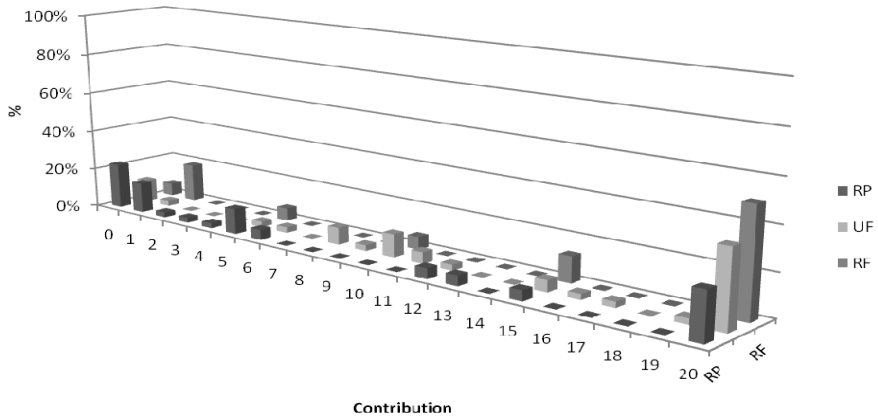

Figure 3. Average quantity of punishment points assigned.

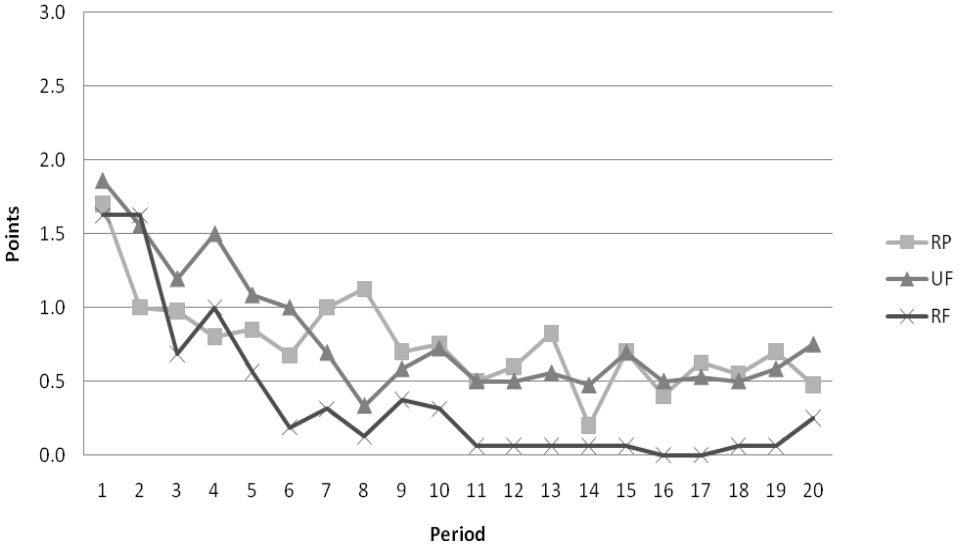

Figure 4. Average quantity of points received as a function of the distance from the average of the group

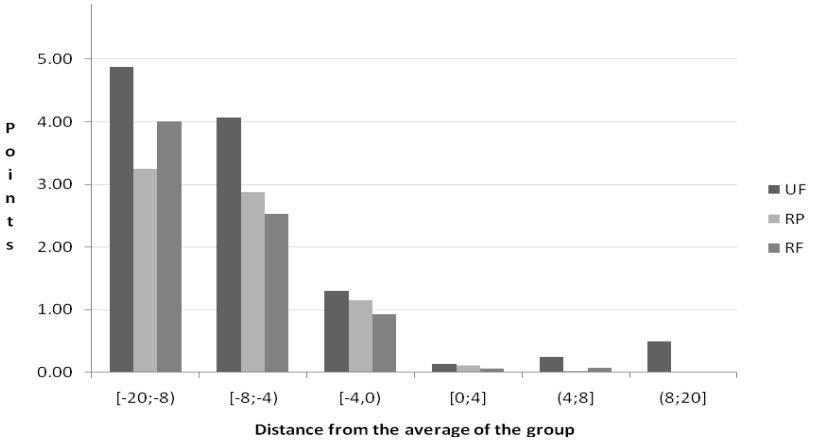

Table 1. Experimental treatments

		Restrictions: Yes	Restrictions: No
Information	Full	RF (4 groups)	UF (9 groups)
	Partial	RP (10 groups)	

Table 2. Cost function

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

Table 3. Average contributions in all periods

	Treatment		
Group	UF	RP	RF
1	13.76	16.06	18.69
	(7.34)	(3.26)	(3.32)
2	18.40	4.05	10.76
	(3.31)	(3.75)	(2.13)
3	4.94	3.43	16.90
	(1.32)	(1.59)	(2.32)
4	11.30	5.44	0.81
	(4.03)	(0.85)	(0.32)
5	12.85	1.74	
	(4.04)	(2.21)	
6	2.84	9.59	
	(2.11)	(1.85)	
7	15.13	19.20	
	(4.77)	(2.20)	
8	7.05	3.70	
	(1.63)	(2.79)	
9	11.11	14.50	
	(2.95)	(2.03)	
10		3.10	
		(1.01)	
Mean	10.82	8.08	11.79

Standard deviations in parentheses.

Table 4. Determinants of contribution

Contribution	**UF + RP + RF**
RP	-3.57* (2.01)
UF	-1.04 (2.75)
Av_first	0.66*** (0. 244)
Constant	-4.61 (6.83)
Log-likelihood	-4995.00
Wald Chi(2)	28.84
N. Of obs.	1840

Legend: the dependent variable takes values from 0 to 20. Av_first: group average contribution in the first period. Controls: gender, age, nationality.

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

Table 5. Average quantity of points given in all periods

	Treatment		
Group	UF	RP	RF
1	0.39	0.93	0.39
	(0.59)	(1.05)	(0.59)
2	0.78	0.78	0.53
	(1.04)	(0.52)	(1.04)
3	0.25	0.25	0.03
	(0.08)	(0.43)	(0.08)
4	0.54	0.54	0.56
	(0.53)	(0.53)	(0.53)
5	0.75	0.85	
	(0.89)	(0.82)	
6	1.65	1.06	
	(1.04)	(1.33)	
7	0.19	0.18	
	(0.25)	(0.46)	
8	1.63	0.98	
	(0.77)	(1.06)	
9	1.54	1.20	
	(0.37)	(0.85)	
10		0.83	
		(0.49)	
Mean	0.81	0.76	0.38

Standard deviations in parenthese

Table 6. Determinants of the quantity of punishment points assigned

Points given	**UF + RP + RF**
RP	-0.07 (0.88)
UF	0.71 (0.83)
Av_first	0.031 (0.10)
Dist_Av	0.68*** (0.05)
Constant	-1.03 (3.00)
Log-likelihood	-1741.32
Wald Chi(2)	212.24
N. Of obs.	1840

Legend: RP dummy variable which assumes value 1 if the treatment is the RP treatment; UF : dummy variable which assumes value 1 if the treatment is the UF treatment the dependent variable takes values from 0 to 20. *Av_first*: group average contribution in the first period. *Dist_Av:* distance from the average contribution of the group. Controls: gender, age, nationality.

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

Table 7. Antisocial punishment in the Baseline

Group	(1) Points given by i to j	(2) Weak a.p. $Contr_j > Contr_i$	(3) % $Contr_j > Contr_i$	(4) Strong a.p. $Contr_j > AV$	(5) % $Contr_j > AV$
1	31	5	16.13%	5	16.13%
2	42	0	0.00%	0	0.00%
3	2	0	0.00%	0	0.00%
4	45	12	26.67%	8	17.78%
5	60	10	16.67%	3	5.00%
6	132	18	13.64%	7	5.30%
7	15	0	0.00%	0	0.00%
8	130	66	50.77%	41	31.54%
9	123	7	5.69%	6	4.88%
Mean	64.44	13.11	14.40%	7.78	8.96%

Table 8: Determinants of the quantity of punishment points received

Received points	**UF**	**RP**	**RF**
Pos_dev_AV	0.03 (0.04)	-0.55*** (0.10)	-0.44*** (0.16)
Abs_Neg_dev_AV	0.68*** (0.82)	0.53*** (0.41)	0.59*** (0.05)
Constant	-3.93* (2.37)	-0.79 (2.73)	-1.72 (2.36)
Log-likelihood	-648.81	-792.13	-226.72
Wald Chi(2)	337.06	231.36	123.60
N. Of obs.	720	800	320

Legend: the dependent variable takes values from 0 to 30. *Pos_dev_AV*: positive difference between a subject's contribution and the group average contribution. Abs_neg_dev_AV: absolute value of the negative difference between a subject's contribution and the group average contribution. Controls: gender, age, nationality.

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

Table 9: Determinants of changes in individual contribution levels (Model 1)

C_t-C_{t-1}	UF	RP	RF
$Point\_rec_{t-1}$	0.83*** (0.15)	1.01*** (0.15)	1.85*** (0.42)
$Deviation\_AV_{t-1}$	-0.79*** (0. 12)	-0.67*** (0.11)	-0.84*** (0.08)
Constant	2.92 (3.42)	-7.22 (6.53)	-5.50 (4.89)
Log-likelihood	-818.50	-675.78	-230.36
Wald Chi(2)	124.78	156.96	93.32
N. Of obs.	648	720	288

Legend: the dependent variable takes values from 0 to 20. $Point\_rec_{t-1}$: number of punishment points received (lagged). $Deviation\_AV_{t-1}$: distance between the subject's contribution and the group average contribution (lagged). Controls: gender, age, nationality.

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

Table 10: Determinants of changes in individual contribution levels (Model 2)

C_t-C_{t-1}	UF	RF
$Point\_rec_{t-1}$	0.32* (0.16)	1.51*** (0.40)
$Deviation\_AV_{t-1}$	-0.43*** (0.13)	-0.08 (0.27)
$Distance\_from\_Highest_{t-1}$	-0.50*** (0. 07)	-0.65*** (0. 19)
Constant	1.14 (3.17)	-5.22 (4.23)
Log-likelihood	-797.09	-224.79
Wald Chi(2)	160.54	108.01
N. Of obs.	648	288

Legend: the dependent variable takes values from 0 to 20. $Point\_rec_{t-1}$: number of punishment points received (lagged). $Deviation\_AV_{t-1}$: distance between the subject's contribution and the group average contribution (lagged). $Distance\_from\_Highest_{t-1}$: distance between the subject's contribution and the highest contribution in the group (lagged). Controls: gender, age, nationality.

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