Threat and Punishment in Public Good Experiments

David Masclet, Charles N. Noussair, Marie-Claire Villeval

Published in *Economic Inquiry* 51(2), 1421–1441, 2013

Threat and Punishment in Public Good Experiments

David Masclet, Charles N. Noussair, and Marie-Claire Villeval

Abstract: Experimental studies of social dilemmas have shown that while the existence of a sanctioning institution improves cooperation within groups, it also has a detrimental impact on group earnings in the short-run. Could the introduction of pre-play threats to punish have enough of a beneficial impact on cooperation, while not incurring the cost associated with actual punishment, so that they increase overall welfare? We report an experiment in which players can issue non-binding threats to punish others based on their contribution levels to a public good. After observing others' actual contributions, they choose their actual punishment level. We find that threats increase the level of contributions significantly. Efficiency is improved, but only in the latter periods. However, the possibility of sanctioning differences between threatened and actual punishment leads to lower threats, cooperation and welfare, restoring them to levels equal to or below the levels attained in the absence of threats.

Keywords: Threats, cheap talk, sanctions, communication, public good, experiment
JEL Classifications: C92, H41, D63

Contact: David Masclet, CNRS, CREM, 7 Place Hoche, 35065 Rennes, France, CIRANO, Montréal. Email: david.masclet@univ-rennes1.fr. Charles N. Noussair, Department of Economics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: C.N.Noussair@uvt.nl. Marie Claire Villeval, University of Lyon, CNRS, GATE, 93, Chemin des Mouilles, 69130, Ecully, France, IZA, Bonn, Germany. E-mail: villeval@gate.cnrs.fr.

Acknowledgements
We thank participants at the International Meetings of the ESA in Washington D.C., USA, the IAREP-SABE in San Diego, USA, at the APET workshop on Behavioral Public Economics in Lyon, France, and seminar participants at the Universities of Innsbruck and Strasbourg for constructive and helpful comments. We thank E. Priour for programming and research assistance. Financial support from the *Agence Nationale de la Recherche* (ANR-08-JCJC-0105-01, "CONFLICT" project) is gratefully acknowledged**.** We thank the associate editor and two anonymous referees for very helpful comments.

1. INTRODUCTION

A large number of experimental economic studies have explored the conflict between individual behavior and collective interest in social dilemmas. One of the principal paradigms employed in this research is the linear Voluntary Contributions Mechanism (VCM) game. In this game, each member of a group of players receives an initial endowment that she may allocate between a private account that returns money only to her, and a group account that benefits all individuals. The payoff structure has the property that each individual has a dominant strategy to allocate all of her endowment to the private account, while the maximum group payoff can only be reached if all members assign their entire endowment to the group account. Laboratory experiments have shown that substantial cooperation, in the form of high assignments to the group account, occurs in the initial periods of play. However, the rate of cooperation decreases as the game is repeated (Isaac *et al*., 1985; Andreoni, 1988; Isaac and Walker, 1988a; Ledyard, 1995).

Two modifications to the game that are known to greatly increase cooperation are to allow pre-play communication (Dawes et al, 1977; Isaac *et al*., 1985; Isaac and Walker, 1988b, 1991; Ostrom *et al*., 1992; Kerr and Kaufman-Gilliland, 1994; Krishnamurthy, 2001; Brosig *et al*., 2003), and to allow players to punish others after contribution decisions are made (Fehr and Gächter, 2000; Masclet *et al*., 2003; Noussair and Tucker, 2005; Bochet *et al*., 2006; Sefton *et al*., 2007; Carpenter, 2007a,b; Egas and Riedl, 2008; Gächter *et al*., 2008). In a common pool resource game, Ostrom *et al.* (1992) compare the impact of various institutions on cooperation and they show that the most effective institution of those they consider is pre-play unstructured communication combined with a voluntary sanctioning institution.

However, while the availability of punishment improves cooperation, the application of punishment is costly to both the sanctioner and the target. In the short-run, the net effect of punishment is to reduce welfare, although punishment increases welfare if the horizon is sufficiently long (Gächter *et al*., 2008). In this paper, we study the effect of permitting explicit, but non-binding, threats to punish. Specifically, we aim to study the impact of a specific form of communication, the issuance of threats, on contributions,

sanctions, and earnings in a VCM game. The conjecture to be evaluated is that threats increase the effectiveness of sanctions. If threats are sufficiently effective in increasing cooperation on their own, then the sanctions need not actually be applied against non-cooperators, and overall welfare might exceed the level achieved in a setting in which no threats could be made. On the other hand, the introduction of explicit threats may crowd out the intrinsic motivation to cooperate. This could be the case, for example, if the threats trigger resentment and result in negative reciprocity from the parties receiving the threats. Such negative reciprocity could take the form of lower contributions or greater punishment assignments. If this occurs, individuals who previously issued strong threats may feel that they must make good on their threats, leading to greater application of sanctions and incursion of costs, and consequently to lower welfare, than in the absence of threats. A third possibility is that the threats have no net effect on welfare. This would be the case, for example, if threats are treated as cheap talk and ignored.

Threats are common in everyday life and often precede sanctions or allow sanctions to be avoided.[1] Parents often use threats to influence children's behavior. Schoolyard bullies issue threats to classmates. Companies sometimes threaten employees to increase productivity. Competing nations threaten each other economically and militarily. Nevertheless, the scientific investigation of the role of threats in human interaction is scant. In experimental economics, we are only aware of a few studies analyzing the behavioral impact of explicit threats to punish (Bochet et al., 2006; Dickinson and Villeval, 2008; Bochet and Putterman, 2009; Li *et al*., 2009).[2]

In the study reported here, we investigate the effect of threats to punish on contributions, punishment, and overall welfare, and also analyze patterns in threats. Our

[1] The situation is somewhat different if one considers exogenous threats such as legal threats (for a recent study on the impact of legal threat campaigns on tax compliance behavior', see Fellner *et al.,* 2009).

[2] In the context of a Voluntary Contribution Mechanism, Bochet *et al*. (2006) compare different forms of communication. They observe that communication has a strong effect on contribution. Adding a punishment option does not raise contributions significantly. In a principal-agent experiment, Dickinson and Villeval (2008) allow the principal to announce threats to monitor and to sanction. They observe both a dominant disciplining effect of threats on effort and a smaller crowding-out effect of threats. Bochet and Putterman (2009) allow people to make non-binding announcements of intended contribution and punishment levels. They find that players increase their contribution announcements in response to others' announcements. Li *et al*. (2009) introduce, in a trust game, threats of sanctions by the trustor before the trustee makes his return decision. Trustees reciprocate less when they face sanction threats.

experimental design has three treatments. The *Baseline* treatment is based on a design used in Fehr and Gächter (2000). In this treatment, the game has two stages. In the first stage, individuals decide, simultaneously, on the portion of their endowment to contribute to the group account. In the second stage, players observe the contribution of each of the other members of their group and simultaneously decide whether and how severely to impose costly punishment on them.

The second treatment is called *Threat*. The Threat treatment is similar to the *Baseline* except that a preliminary stage is included, in which players announce a threat to punish. They do so by submitting a threat schedule. They must specify a function, which indicates how much they threaten to punish other members of their groups, for each level they could contribute. The schedule can condition only on contribution level, and is uniform across all potential recipients.

The third treatment, called the *Second Order* treatment, differs from the Threat treatment in that a fourth stage is added to the game. In this final stage, players are informed of other group members' threats and the sanctions they assigned, so that they can observe the extent to which other individuals carried out their threats. The players can then assign additional punishment, potentially punishing those who did not carry out their threats. As a consequence, we may conjecture that this possibility of second order punishment reduces the difference between threatened and realized punishment levels, since differences can be punished. This may occur either by reducing the level of threats or by increasing the severity of punishment assignments.

Our paper is most closely related to the studies of Bochet *et al*. (2006) and Bochet and Putterman (2009). Bochet *et al*. (2006) compare different forms of communication: numerical, face-to-face, and a computerized chat, in the context of a Voluntary Contribution Mechanism, in conjunction with punishment. In the absence of punishment, face-to-face and chat increase cooperation, but numerical communication does not. Adding punishment when communication is already possible does not affect cooperation. Bochet and Putterman (2009) allow people to make non-binding announcements of potential contributions and punishments, which in some treatments can be labeled as

promises. Punishment announcements are directed toward those who contribute less than the average. Observing a punishment announcement causes contribution announcements to increase. Players are punished for contributing less than their announced level, especially when the announcement is a promise. Overall, promises increase cooperation and earnings in a setting in which punishment is possible.

Our design differs from that from these previous studies. In our design participants cannot make non-binding announcements of contribution levels, but only announcements of potential punishment levels. The schedule of threats is uniform for all other group members and cannot be conditioned on the contribution profile of the whole group. Furthermore, in contrast to these previous studies, participants are not allowed to revise their messages in response to others' messages. Finally, in our design, all players can observe how much punishment is threatened for each possible contribution level. In contrast, in the previous studies mentioned above, participants could observe threats of punishment only for the announced intended contribution levels.

Our results show that communication, in the form of threats, increases contributions, even though threats are cheap talk. Threat levels are positively correlated with, but typically greatly overstate, the subsequent sanctions. Players punish a given contribution more heavily in the Threat than in the Baseline treatment. Initially, the benefit to welfare of the higher contributions and the cost of the greater punishment offset, so that threats do not increase efficiency in the short term, though there is a modest improvement in welfare after players have experienced many periods of play. Permitting punishment of differences between threats and actual sanctions has the effect of reducing the difference between threats and sanctions through a reduction in the intensity of threats. Failure to carry out threats draws punishment. However, on the whole, cooperation, punishment, and therefore welfare, are reduced to levels similar to the Baseline treatment. The main findings are robust to a change in the cost that individuals must pay to apply punishment.

The remainder of the paper is organized as follows. In section 2, we describe the experiment. Section 3 presents the results and section 4 consists of a brief discussion.

2. THE EXPERIMENT

The experiment consisted of 16 sessions conducted at the LABEX facility of the Center for Research in Economics and Management (CREM), at the University of Rennes I, located in Rennes, France. The 200 participants were recruited from various undergraduate courses. No subject participated in more than one session. The experiment was computerized using the Ztree software package (Fischbacher, 2007), and conducted in French. On average, participants earned 14 Euros, including a €3 show-up fee. Table 1 provides some information about the individual sessions. Participants interacted during 20 periods under a partner matching protocol.[3]

2.1. The Baseline Treatment

Our experiment has three treatments, called *Baseline*, *Threat*, and *Second Order*. As described below, each treatment is conducted under both a Low (LE) and a High (HE) Effectiveness condition, though our analysis will focus predominantly on the data from the HE condition. A session conducted under any of the treatments consists of a series of 20 periods. Each period of the Baseline treatment has two stages. At the beginning of stage one, each member of a group of four players receives an endowment of 20 ECU, an experimental currency convertible to Euros, to allocate between a private account and a group account. No player can observe any other player's contribution decision before he makes his own choice. Each ECU that any group member allocates to the group account yields 0.4 ECU to each member of the group. The payoff of subject *i*, at the end of the first stage, π_i^1, equals:

$$\pi_i^1 = (20 - c_i) + 0.4\sum_{j=1}^{4} c_j \qquad (1)$$

where c_i is player *i*' s contribution to the group account. The more ECU an individual allocates to the group account, the lower her own but the greater the group's total

[3] To avoid reputation effects across periods, participants were associated with a letter of the alphabet, A,..,D that was randomly changed after each period. An individual's activity was displayed in a different position on other group members screens in different periods. This made it impossible for an individual to track another player's behavior from period to period.

earnings. For this reason, allocations to the group account are referred to as contributions, and higher contributions can be interpreted as greater cooperation.

Each participant is then informed of her first-stage payoff, the total contribution of the group, and the individual contribution of each of the three other members of her group. In stage two, she has an opportunity to assign punishment points to each of the other members of her group. No player could observe any other's punishment decision at the time she made her choices. Each individual assignment was required to be in the range from 0 to 10. Under the High Effectiveness condition, each point assigned costs one ECU to the punisher and two ECU to her target. Under the Low Effectiveness condition, each point assigned costs one ECU to the punisher and one ECU to the target. Therefore, player *i*'s payoff after the second stage is given by:

$$\pi_i^2 = \pi_i^1 - \varepsilon \sum_{j \neq i} p_j^{i2} - \sum_{j \neq i} p_i^{j2} \tag{2}$$

where p_i^{j2} is the number of points *i* assigns to *j* in the second stage. The parameter ε equals 2 in the HE and 1 in the LE condition. Previous research shows that the inclusion of a punishment opportunity with ε = 2 leads to higher cooperation in the conditions of our Baseline treatment relative to a setting with no punishment, while ε = 1 fails to increase cooperation (Nikiforakis and Normann, 2008).

2.2. The Threat and Second Order Treatments

The Threat treatment is identical to the Baseline except that a preliminary stage with structured communication is included. In this additional stage, which we refer to as stage zero, the players were required to simultaneously announce a hypothetical punishment level in the range of 0 to 10 for each possible contribution level that a member of their group could make in stage one. This announcement was non-binding, but was communicated to the relevant parties. In this paper, for clearer exposition we refer to these hypothetical punishment points as 'threat points', to avoid any confusion with the actual punishment points distributed later in the period.

The purpose of this stage was described to the players in the following terms:

"You announce the number of points you would like to assign to each other group member for each possible contribution level (between 0 and 20 ECU) to the project in the second stage. The number of points you announce for a group member indicates your degree of disapproval for each contribution level (from 10 points for the highest disapproval to 0 point for no disapproval)."[4]

After threat points are assigned, but before contribution decisions are made in stage one, the players are informed of the total number of threat points the three other members of his group have assigned for each possible contribution level. That is, denoting $t_j(c)$ as the function indicating how many threat points that player j assigns for each contribution level c, each player i learns of $\sum_j t_j(c)$ for all possible contribution levels c. This means that players could indicate how much punishment they would give for each possible contribution level. However, they could not issue different threats to different individuals. This means that the threat could not condition on the specific profile of individual contributions, preventing punishment based on, for example, the recipient's contribution compared to the group average, or whether the recipient was the lowest contributor in the group.

Stages 1 and 2 proceed in the same manner and have the same payoff structure as the Baseline treatment. It is common knowledge, from the public reading of the instructions, that the number of punishment points assigned is not required to match the number of threat points the player announced previously (see the instructions reprinted in the *online* appendix).

The Second Order treatment is identical to the Threat treatment except that an additional stage[5] is included at the end of each period. This final stage consists of an additional round of sanctions. At the beginning of this final stage, each player i is informed of the number of punishment points each other player j has directed toward

[4] In all treatments, we chose to use the term "disapproval" to facilitate the understanding of the participants in the pre-play communication stage and to make sure that they were aware of the precise purpose of the threat points (see Masclet *et al*., 2003). We acknowledge that the use of this term may have affected the responses of those who assigned and received the threat points. However, because behavior in our Baseline treatment, in which we also use the term "disapproval points," does not differ qualitatively from previous studies, the impact of our terminology is unlikely to have had a drastic impact on behavior.

[5] In the instructions distributed to the subjects for the Threat treatment, the stage in which threats are submitted is called stage one, the contribution stage is called stage two, and the punishment stage is called stage three. In the Second Order treatment, the same designations are used as in the Threat treatment, and the second round of sanctions is referred to as stage four.

every player $k \neq i$, as well as the threat that j had specified against k's actual contribution level.[6] This means that players can observe any difference between the threats announced in stage zero and the actual punishment assigned in stage two, except for those assigned to him. This is an important difference from the Threat treatment, in which participants were not informed about the actual sanctions assigned to other group members, and therefore could not observe a possible discrepancy between threats and actual sanctions. The fact that participants cannot observe who punished them precludes retaliation. Nevertheless, individuals may form beliefs about who sanctioned them based on who sanctioned others, and may attempt to counter-punish based on this information.

Then, each player can assign additional punishment points. The cost of these points is the same as for punishment points assigned in the previous stage. Individuals are not informed about who sanctioned them and by how much, in either punishment stage. They also do not observe the second-order punishment decisions of other group members.[7] The final payoff in a period, for individual i in the Second Order Treatment, is:

$$\pi_i^2 = \pi_i^1 - \varepsilon\left(\sum_{j\neq i} p_j^{i2} + \sum_{j\neq i} p_j^{i3}\right) - \left(\sum_{j\neq i} p_i^{j2} + \sum_{j\neq i} p_i^{j3}\right) \qquad (3)$$

A key feature of the design to bear in mind is that the information that is available on which to base punishment differs between the three treatments. In the Baseline treatment, individuals can punish on the basis of their own and others' contribution behavior. In the Threat treatment, they can punish based on their own and others´ contributions, as well as on the threats that they and others have made. In the Second Order treatment, they can

[6] We recognize that having a second punishment opportunity, in which the only new information available is the discrepancy between others' threats and their actual punishments, is somewhat artificial. Providing more information to the participants at this stage, however, might have introduced additional motives to apply second-round punishment, and may have made it more difficult to associate second-round punishment with subsequent threat and punishment decisions.

[7] Nikiforakis (2008) allows players to observe individual punishment behavior, and makes reprisals possible. Reprisal opportunities tend to offset the positive effect of punishment on contributions. Other studies have investigated the effect of allowing subjects to punish second order free riding (i.e. punish those who failed to punish low contributors, Cinyabuguma *et al*., 2006, Denant-Boemont *et al*., 2007). These experiments suggest that allowing sanction enforcement causes a modest increase in contributions.

punish for the same motives as in the Threat treatment, but also on the basis of the difference between threatened and actual punishment assigned or received.[8]

In all treatments, assuming that players maximize their own earnings, the subgame perfect equilibrium is to not contribute at all to the public good and not to punish at any decision node. The marginal per capita return of the public good is always lower than the marginal return of keeping one's own endowment for oneself. In contrast, the socially optimal behavior is to contribute one's full endowment to the public good, since $0.4*n > 1$. In the treatments with threats, any profile of threats is compatible with the subgame perfect equilibrium, since threats are cheap talk and the equilibrium is unique. No punishment is observed in equilibrium in any treatment since assigning punishment always reduces the payoff of the punisher.

[Table 1 about here]

3. RESULTS

This section is organized as follows. In section 3.1, we consider patterns in the assignment of threats. We then turn to the relationship between threats assigned and subsequent punishment. In section 3.2, we study the differences in contributions and earnings between treatments. The analysis in sections 3.1 and 3.2 concentrates on the HE condition. We focus on HE because it is a condition in which punishment is known to work, in the sense that it typically induces a positive effect on contributions under Baseline conditions (Nikiforakis and Normann, 2009). In section 3.3, we consider whether the results are similar in the LE condition and establish that many of the patterns observed in HE are robust to the difference in punishment effectiveness.

3.1. Threats and sanctions

3.1.1. Assignment of threats

[8] It is worth noting that players may also sanction in this additional stage of the Second Order treatment for several other motives including blind retaliation, use of a second opportunity to punish low contributors, punishment of those who failed to punish low contributors in the preceding stage and punishment of those who punished high contributors. We attempt to control for all of these motives in our data analysis.

Figure 1 displays the average threat assigned for each possible contribution level in each of the treatments in the High Effectiveness condition. The figure shows that threats are widely employed. In 83.75% of instances in the Threat treatment (469 observations out of 560), the threat schedule a participant submits contains a threat to punish at least one contribution level. By this criterion, threats are made in 87.36% of possible instances (629 observations out of 720) in the Second Order treatment.

The figure also reveals that individuals make less severe threats against higher contributions, and for all possible contribution levels, the average threat is higher in the Threat than in the Second Order treatment. The average threat from one individual to another is 7.34 and 6.68 for a contribution level equal to zero in the Threat and Second Order treatments, respectively. On average, threats of 0.66 and 0.33 are made for the highest possible contribution of 20. In 11.61% of the observations in the Threat treatment, and 6.53% in the Second Order treatment, threat points are directed at even the highest possible contribution. Threats are on average considerably more severe for contributions just below the maximum, however. Averages of 3.77 and 2.34 threat points are assigned for a contribution of 19 in the Threat and Second Order treatments, respectively. 51.96% of the players in the Threat treatment, and 38.47% of those in the Second Order treatment, threaten to punish a contribution of 19. Our findings regarding threat decisions are summarized in Result 1.

[Figure 1 and Table 2 about here]

RESULT 1: *Threats are widely employed, even against those making high potential contributions. Threats are more severe against lower contributions. For all contribution levels, threats are less severe in the Second Order treatment than in the Threat treatment. Threat severity increases over time, with the exception of the last period.*

Support for Result 1: Table 2 contains the estimates of five random-effects Tobit models, in which the dependent variable is the number of threat points that player i assigns to player j (for $j \neq i$) for a given level of contribution c. In models (2) to (5), c takes

the following values: c = 0, 10, 15 and 19.[9] In all of the regressions, the independent variables include a dummy variable for the Second Order treatment (so that the Threat treatment is the reference category), a time trend, and a dummy variable for the final period.

The estimates show that significantly fewer threat points are assigned in the Second Order than in the Threat treatment for any positive contribution level. The significant time trend for all contribution levels indicates that threats tend to escalate over time, controlling for the threatened player's contribution level. This might be interpreted as reflecting a process whereby contributions increase over time, eventually increasing the contribution norm. More severe punishments are assigned for given contribution levels, as they represent lower and lower levels relative to the norm.

In another Tobit regression (see Table A1 in the online appendix), the dependent variable is the contribution threshold above which the player no longer threatens to punish. The independent variables are the same as in the regressions of Table 2. The results indicate that the contribution threshold, above which people cease threatening others, does not differ across treatments (coeff. = -0.497, S.E. = 1.165) and increases over time (coeff. = 0.085***, S.E. = 0.018) (N=3840, left censored observations=551, right-censored observations = 336; log-likelihood = -10331.63).

3.1.2. The relationship between threats and first order punishment

We have seen that heavy threats are issued. We now consider the consistency of threats with subsequent punishment decisions. Our findings are reported in Result 2.

RESULT 2. *Actual sanctions are much less severe than those that are threatened. Threatened sanctions for realized contribution levels are, nevertheless, positively correlated with subsequent sanctions. The severity of sanctions decreases over time, while the severity of threats increases over time.*

[9] We did not run a similar Tobit estimate for a contribution level of 20 because there are too many values of 0 for the dependent variable (91.25% of all observations).

Support for Result 2: On average, participants assign 0.423 punishment points in stage two of the Baseline treatment (S.D. = 1.42) 0.61 points in the Threat treatment (S.D. = 1.76), and 0.45 in the Second Order treatment (S.D. = 1.48). The number of punishment points is higher under the Threat treatment compared to the Baseline treatment, although this difference is not significant. Mann-Whitney pairwise tests, with each group's decision as an observation, conclude that there is no difference in punishment levels between the Threat and the Baseline treatments (z =-0.384, p= 0.701), between the Second Order and the Baseline treatments (z =-0.795, p= 0.426), or between the Second Order and the Threat treatments (z = 0.476, p =0.633).

Figure 2 displays the average number of threat points issued for the *actual* subsequent realized contribution levels (the threat schedules evaluated at actual subsequent contributions), and the actual number of punishment points assigned in the second stage of both the Threat and the Second Order treatments. For comparison, Figure 2 also displays the average number of points assigned in the Baseline treatment. These are displayed as a function of the difference between the target's contribution and the average group contribution (excluding j's contribution), in the High Effectiveness condition.[10] Figure 2 shows that punishers react strongly to negative deviations from the average contribution. For the purpose of comparison, the threat points are also shown in the figure. The figure suggests that the intensity of the threat level is a good indicator of subsequent punishment decisions, in the sense that threats and punishment are correlated. However, actual sanctions administered are far less severe than those that were threatened. For example, a subject who contributes between 14 and 20 units less than the group average in the Threat treatment receives on average 8.03 threat points but 3.92 punishment points. Lastly, Figure 2 indicates that for all intervals, more punishment points are assigned in the Threat treatment than in the Baseline.

[10] Figure 2 should be interpreted while taking into account that differences from the average contribution were not yet observable when the threats were issued. Figures A1 and A2, in the online appendix, display the threatened and actual punishment as functions of absolute contribution levels of the recipient. However, Figure 2 is arguably more relevant than these two additional figures because actual sanction assignments tend to depend more on deviations from the average than on absolute contribution level.

[Figure 2 about here]

The left panel of Table 3 reports the estimates of three random-effects Tobit models. The dependent variable is the number of punishment points that *i* assigns to *j* in the punishment stage of period *t* following the contribution stage. The first two models use the pooled data from the three treatments, while the third model uses the pooled data from the Threat and Second Order treatments. The independent variables include dummy variables for the treatment in effect, the average amount contributed by the group (excluding *j*'s contribution), the differences between *j*'s and the average contribution in the group conditional on *j* contributing less or more than the group average, a time trend, and a dummy variable for the final period. In the third model, the regressors also include the threat assigned by *i* for *j*'s actual contribution. In addition, a dummy variable entitled "Anti-Social Threatener' indicates whether *i* has made a threat for the highest possible contribution.

[Table 3 about here]

Table 3 indicates that participants receive more punishment, the less they have contributed relative to their group's average. This pattern is in agreement with previous studies (for example see Fehr and Gächter, 2000; Masclet *et al*., 2003; Nikiforakis, and Normann. 2008; Nikiforakis, 2008; Bochet et al., 2006; Bochet and Putterman, 2009). Model (2) shows that, controlling for the differences between the target's and the average contribution in the group, participants punish slightly more in the Threat treatment than in the Baseline. Estimated equation (3) in Table 3 shows that the stronger the prior threat, the more punishment points are assigned. Furthermore, the participants who threaten to punish the highest contribution level are more willing to sanction others.[11] Thus, the severity of a threat is an indicator, albeit a biased one, of subsequent sanctioning decisions.

[11] In an additional regression (see Table A2 in the online appendix), we have tested, using a random-effects probit model, whether being an anti-social threatener is an indicator of the probability of being an anti-social punisher (i.e. of punishing those who contributed more than the punisher). The coefficient of this variable is highly significant (at the 1% level), indicating that anti-social threateners are also relatively likely to be anti-social punishers.

3.1.3. Threats and second order punishment

In the Second-Order treatment, players may observe and punish differences between threatened and actual stage two sanctions. As we indicate in result three, empty threats, those that exceed the eventual punishment applied, are indeed sanctioned.

RESULT 3.*Individuals sanction those who fail to carry out their threats.*

Support for Result 3. Consider the four regressions reported in the right panel of Table 3. The dependent variable is the number of punishment points that *i* assigns to player *j* in the second round of sanctions in period *t*. With these estimations, we investigate to what extent second order punishment is used to sanction those who fail to carry out their threats. We attempt to control for other motives that might also induce people to punish at this stage. These other motives include *i*) use of this second opportunity to punish low contributors, *ii*) blind retaliation for punishment received in the previous stage (blind counter-punishment), *iii*) punishment of those who failed to punish low contributors in the preceding stage (sanction enforcement), and *iv*) punishment of those who punished high contributors. In model (4), the independent variables are the average group contribution (excluding *j*'s contribution) and the absolute values of positive, as well as of negative, differences between *j*'s contribution and the average contribution of others. These variables capture the punishment of low (high) contributors. The specification also includes, as dependent variables, the average threat made by *j* to players *k* other than *i* for their actual contribution levels, and the number of punishment points *j* actually assigned to them. A continuous variable, "how much more j threatens than he punishes", captures the impact of player *j* punishing less than he threatened.[12] A dummy variable registers whether player *i* has been punished or not in the first round of sanctions. There is also an interaction term measuring whether player *i* has been punished in the first round of punishment times the number of punishment points *j* assigned to others. These variables aim at capturing intended counter-punishment. Indeed, as mentioned above, although participants could not observe individual

[12] This variable takes the value of the difference between threats and punishment assigned by *j* in case *j* has assigned more threat points than punishment points, and 0 otherwise. To avoid perfect collinearity with the variables "j's average threat", and "j's average punishment in first round", the symmetric variable ".how much less j threatens than he punishes" is not included in the model.

assignments of punishment directed toward themselves, they may have formed beliefs about who sanctioned them based on who sanctioned others, and may seek to retaliate based on this information.

Model (5) includes the same variables as model (4) plus the positive and negative differences between the number of punishment points assigned by *j* to other players (excluding *i*) and the average assignment to these players. These variables aim at controlling for sanction enforcement. The positive and negative differences are written as:

$$\max\left\{\sum_{k\neq i} p_j^{k1t} - \left(\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k1t}\right)/2, 0\right\}$$

and

$$\max\left\{0, \left(\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k1t}\right)/2 - \sum_{k\neq i} p_j^{k1t}\right\}$$

respectively. Model (6) also includes a dummy variable indicating whether player *j* was a perverse punisher, that is, whether player *j* has punished an individual who contributed more than the group average (see Bochet et al., 2006 and Cinyabuguma et al., 2006). Alternatively, in model (7) we have replaced this variable by an indicator of whether *j* is an anti-social punisher, that is, whether player *j* punished someone who contributed more than he did himself in the current period (see Herrmann et al., 2008).

The four estimations show that a subject is more likely to be punished in the second round of punishment, the fewer punishment points he assigned compared to the quantity he threatened to assign. Empty threats are punished. However, there is also evidence of other motives to punish in this second punishment stage. Low contributions are punished again in this stage, as indicated by the significant coefficient associated with the negative difference between *j*'s contribution and the group average. Moreover, a subject who has been punished in the first punishment stage is more likely to punish in the second punishment stage, even though he does not know who directed the punishment at him previously. The significant coefficient of the number of punishment points assigned by player *j* may indicate an attempt to counterpunish on the part of *i*, who might

interpret a large assignment of punishment to others as an indication that *j* is likely to have been the one who punished *i*.[13] Those who make anti-social threats are more likely to punish in the final punishment stage. The coefficients of the variables "*j* is anti-social punisher" and "*j* is perverse punisher" are not significant, indicating that in our setting, second order punishment is not used to deter anti-social or perverse punishment. Finally the fact that the variables controlling for positive and negative differences between the number of points assigned by *j* to other players and the average assignment to these players are not significant, indicates the absence of sanctions against those who failed to punish low contributors, once the effect of deviations from threatened punishments is taken into account.

3.1.4. Implications of receiving second order punishment on subsequent threats

If failure to carry out threats is punished, participants may react by reducing their threats. We observe that this is indeed the case, as argued in Result 4.

RESULT 4. *Threat behavior responds to the punishment of empty threats. In the Second Order treatment, participants who punish less than they threaten to, and who are subsequently punished, decrease their threats in the next period.*

Support for Result 4. We have estimated a model of the determinants of changes in the threats made between periods *t* and *t+1* (See Table A3 in the online appendix). This model is estimated separately for the participants who threatened more and those who threatened less than they actually punished in period *t*. The independent variables consist of both the difference between the number of threat points and the actual sanctions assigned by player *i* to the other members of his group after being informed of their

[13] In estimation (7), the marginal effects of the variables "sanctions *i* received in first round", "*j*'s average punishment in first round" and of the interaction term "sanctions *i* received in first round**j*'s average punishment in first round" amount to 0.164, 0.110 and -0.046, respectively. This means that those who have been punished in the first punishment stage assign on average 16.4 per cent more of punishment points in the second stage of punishment. To obtain the interaction effect for having been punished in the first round of punishment and observing that *j* has assigned punishment points to others, add the marginal effect 0.164 (sanctions *i* receives in first round) to the coefficient 0.110 (*j*'s average punishment in first round) and the coefficient - 0.046 (sanctions *i* received in first round**j*'s average punishment in first round). This gives 0.228, indicating that those who have been punished in the first punishment stage and observe that *j* has assigned punishment points to others increases their punishment in second round of sanction by 22.8 per cent.

contribution levels, and the total number of punishment points received by *i* in the final stage of period *t*.

The estimates show that individuals who assigned more threat points than first-round punishment points in period *t* respond to second-round sanctions by revising downward the number of threat points they assign in the following period (coeff. = -0.170, p = 0.028). Moreover, the greater the difference in period *t*, the more they revise downward (coeff. = -0.452, $p< 0.001$). No such adjustment is observed for those who punished either equally or more severely than their threats ($p = 0.570$ and $p = 0.814$, respectively).

3.2. Contributions and earnings

3.2.1. The effect of threats and sanctions on contributions

We now turn to treatment differences in contribution levels to examine whether threats influence cooperation. Figure 3 displays the time path of individual contributions by period, averaged across groups, in the three treatments, under the High Effectiveness condition. Our observations regarding contribution levels are described as Result 5.

[Figure 3 about here]

RESULT 5: *The possibility of issuing threats increases cooperation. In the Threat treatment, average contributions are greater than in Baseline. However, the additional sanctioning possibilities available in the Second Order treatment reduce cooperation to a level equal to that in the Baseline treatment.*

Support for Result 5: As shown in Figure 3, non-binding threats of punishment increase average contributions in the High Effectiveness condition. The average contribution levels are highest in the Threat treatment (mean = 18.19 ECU per individual, S.D. = 3.32), followed by the Baseline (16.05 ECU, S.D. = 5.00), and by the Second Order treatment (15.95 ECU, S.D. = 4.90). Two-tailed Mann-Whitney pairwise tests, with each group average contribution over the session as an independent observation, indicate that the difference between the Baseline and Threat treatments (p = 0.06), as well as the difference between the Threat and the Second Order treatments (p = 0.08), are borderline

significant. In contrast, there is no significant difference between the Baseline and the Second Order treatments (p= 0.750).

We have estimated several regressions in which the dependent variable is the player's contribution. Table 4 reports the results of these estimations. The independent variables include dummy variables for treatment, a time trend, and a dummy variable for the final period. In regressions 1, 3, and 7, the data from all the treatments are pooled together, and the reference category is Baseline. The independent variables also include the number of threat points received from the other three group members averaged over all possible contribution levels, the total number of threat points received for the highest possible contribution of 20 and the ratio of the threat for a contribution of 20 and the average threat assigned for contributions of less than 20. They also include the threshold at which the subject no longer makes threats, and a dummy variable indicating whether the subject threatens others making the highest possible contribution.

[Table 4 about here]

Table 4 shows that the participants contribute more in the Threat treatment than in the Baseline (see (1) and (3)). On average individuals invest 2.14 ECU more in the group account in the Threat treatment (regression (1)). Participating in the Threat treatment makes a significantly positive difference on contributions from the very beginning of the game, as indicated by regression (7). In contrast, controlling for the threats received, individuals contribute significantly less (-1.91 ECU) in the Second Order treatment than in the Threat treatment (regression (2)). The estimation of the tobit models confirms these findings.

Models (2) and (4) also show that receiving other players' threats increases cooperation significantly. In contrast, controlling for the general impact of threats, models (5) and (6) reveal that participants react to anti-social threats (those directed towards the highest possible contribution) by reducing their contribution. This may result from the fact that individuals have less incentive to raise their contribution if they know that they will be punished in any case. This may also result from the fact that assigning points for the maximum possible contribution level signals that some members in the

group will likely contribute at a lower level. Indeed, those who assign threats for the highest possible contribution of 20 ECU cooperate significantly less, which is consistent with previous findings about perverse or anti-social punishment (see Bochet et al., 2006; Cinyabuguma, Page and Putterman, 2006; Herrmann et al. 2008). Contributions increase significantly over time (except in the final period).

The number of sanctions received in the previous period has not been included in these regressions to avoid autocorrelation. To measure their impact, we have estimated the magnitude of some influences on changes in individual contributions between periods *t* and *t*+1 in separate random-effects GLS regressions (See Table A4 in the online appendix). We conducted the estimations separately for the participants who contribute less than the group average (designated as low contributors), and for those who contribute more than the average (high contributors), in period *t* (N = 457 and 1291, resp.; R^2 = 0.429 and 0.081, resp.). We also include terms for interactions between the punishment received and treatment, as well as for the difference between *i*'s own and the others' average contributions.

The estimates show that, while sanctions increase subsequent contributions of low contributors (coeff. = 0.316, p = 0.001), they have no impact on the behavior of high contributors (p = 0.635).[14] The impact of the first round of punishment on subsequent contributions is similar in the Threat and the Second Order treatments as in Baseline (p = 0.763 and p = 0.487 for low contributors, p = 0.881 and p= 0.374 for high contributors, respectively). Similar regressions for the second round of punishment in the Second Order treatment indicate that sanctions received in the final punishment stage have no impact on subsequent contributions (low contributors: p = 0.178, N = 198, R^2 = 0.546; high contributors: p = 0.191, N=284, R^2 = 0.105), suggesting that receiving such sanctions is not interpreted as punishment for a low contribution.

3.2.2. The effect of threats and sanctions on earnings

[14] This last finding differs from one reported in Masclet et al. (2003), Cinyabuguma et al. (2006), Ones and Putterman (2007), and Page *et al*. (2008). They find that punished high contributors reduce their contribution on average.

As suggested earlier, if threats are effective in inducing greater cooperation, then the sanctions may not need to actually be implemented. Such a pattern would minimize the detrimental effects of punishment on efficiency and result in an improvement in overall welfare compared to a setting in which no threats can be sent. The data supports this hypothesis, but only partially, as summarized in Result 6.

RESULT 6. *Threats do not increase earnings if all periods are considered. They increase earnings in the latter periods of interaction. The ability to punish discrepancies between threats and sanction assignments reduces earnings. Earnings in the Second Order treatment are below the Baseline treatment.*

Support for Result 6. The mean payoff after the contribution stage amounts to 29.63 ECU in the Baseline treatment (S.D. = 4.95), 30.92 in the Threat treatment (S.D. = 3.45), and 29.57 ECU in the Second Order treatment (S.D. = 5.20). However, the positive effect of threats on cooperation is partly offset by the cost of sanctions. The direct cost of punishment can be easily measured by comparing the average payoff after stage one and at the end of the period, in each treatment. The final payoffs amount to 25.84 ECU in the Baseline (S.D. = 8.31; this corresponds to 87.21% of the stage one payoff), 25.47 ECU in the Threat treatment (S.D. = 9.31; 82.37% of the stage one payoff), and 23.20 ECU in the Second Order treatment (S.D. = 11.17; 78.46% of stage one payoff). The relatively higher cost in the Threat treatment compared to the Baseline results from greater point assignments under the Threat treatment (see Table 3 and Figure 2). The relatively low payoff in the Second Order treatment results both from a smaller impact of threats on contributions, as well as from higher costs of punishment, due to the existence of two punishment stages.

Figure 4 displays the difference in the average group payoff between the Threat (Second Order) treatment and the Baseline treatment, and normalizes this difference by the average group payoff of the Baseline treatment in the same periods. It illustrates the evolution of the relative payoff gain/loss in the Threat and Second Order treatments over time, respectively. Figure 4 shows that the Threat treatment succeeds in generating greater earnings than the Baseline treatment only in the late periods. The final (total end-

of-period) payoffs are 25.92 ECU in the Second Order treatment (S.D. = 7.82), 27.02 ECU in the Baseline (S.D. = 7.36) and 27.83 (S.D. 7.20) in the Threat treatment in the last 10 periods of the sessions. As shown by Figure 4, the relative payoff gain of the threat treatment is higher if one considers the last five periods of the game. It amounts on average to 10 percent compared to the Baseline. The final payoffs in periods 15-20 average 29.31 ECU in the Threat treatment (S.D. = 4.97) and 26.99 ECU in the Baseline (S.D. = 7.11). While Gächter et al. (2008) show that the benefits of sanctions may increase over a long-term interaction, the total duration of our experiment is not long enough to confirm this tendency. In contrast, the Second Order treatment induces a relative loss compared to the Baseline treatment throughout the session. The final payoff averages 26.13 ECU in the five last periods of the Second Order treatment (S.D. = 7.00).

Table 5 reports the estimations of three models, in which the dependent variable is the stage one payoff (model (1)), or the final payoff (models (2) to (5)). The independent variables include treatment and a dummy variable for the last 10 periods of a session. Lastly, a dummy variable interacting the Threat (Second Order) treatments and the last ten periods are also included in the estimates. Own contribution is also included as an independent variable.

[Table 5 and Figure 4 about here]

The dummy variable for the last 10 periods is not significant in model (1) whereas it is positive and significant in models (2) and (3). This confirms the fact that final payoffs are significantly higher in the last periods of the game as fewer sanctions are assigned over time. The coefficient associated with the variable "own contribution" is positive and highly significant in model (1) but negative in models (2) to (5), suggesting that free riding becomes unprofitable in the presence of punishment. The Threat treatment induces significantly higher stage one payoffs than the Baseline (model (1)). No effect of the Threat treatment is found on welfare in terms of end-of-period payoffs except if one considers the last five periods of the game (see model (5)). In contrast, in the Second Order treatment, payoffs do not differ from the Baseline after stage one (see model 1), but are significantly lower at the end of the period (see model 2).

3.3. The impact of punishment effectiveness

In this subsection, we consider the data from the Low Effectiveness condition. As under HE, in the LE condition the average individual contributions are the highest in the Threat treatment (11.51, S.D. = 2.13), followed by the Baseline treatment (10.07, S.D. = 6.08), and by the Second Order treatment (8.86, S.D. = 5.39). Figure 5 displays the behavior of contributions over time. It shows that the effect of threats on contributions is less persistent over time than in the HE condition. Our findings are summarized in Result 7.

[*Figure 5 about here*]

RESULT 7: *The number of threat points assigned is similar in the LE and the HE conditions. Under LE as under HE, the Threat treatment has a positive effect on the average contributions compared to the Baseline treatment. This effect is less persistent over time in the LE condition than in HE. Threats do not increase payoffs in LE. Earnings are lower in the LE than in the HE condition.*

Support for Result 7: GLS regressions indicate that the contribution threshold at which a subject no longer assigns threat points is similar in the LE and HE conditions of the Threat treatment (N = 1280; $p = 0.651$) and of the Second Order treatment (N = 1440; $p = 0.274$). The number of threat points assigned is the same in both treatments for every level of contribution ($p > 0.100$), except that the average threat against the maximum contribution are higher in the LE than in the HE condition of the Second Order treatment (N =1440; $p = 0.037$). However, because in the HE treatment, each point assigned results in twice the reduction in earnings, the threatened reduction is greater in HE.

A Mann-Whitney pairwise test comparing average contributions in the Threat and the Baseline treatments in the LE condition indicates that people contribute significantly more in the Threat treatment than in the Baseline in the first ten periods (p=0.070). No significant difference is found between these treatments after period 10. While the average contribution is higher in the Threat than in the Second Order treatment in the first ten periods ($p = 0.050$), no significant difference is found in the second half of the game.

A Mann-Whitney test comparing contributions in the Baseline treatment in HE (averaging 16.05 ECU) and in LE (averaging 10.07 ECU) indicates that people contribute significantly more in the HE condition (p = 0.007). Similar results are obtained when comparing contributions in the Threat treatment in the LE condition (11.51 ECU) and the HE condition (18.19 ECU) (p= 0.053), and when comparing contributions in the Second Order treatment in the LE (8.86 ECU) and HE conditions (15.95 ECU) (p= 0.012).

There is no difference in final period earnings between the Baseline treatment (22.42) and the Threat treatment (22.97, p = 0.965), while payoffs are significantly lower in the Second Order treatment than in both the Baseline (16.29; p = 0.015) and the Threat treatments (p = 0.024). Final earnings are lower in LE than in HE for the Baseline (22.42 and 25.84 ECU; p = 0.101) and the Second Order treatment (16.29 and 23.20 ECU; p = 0.038). In the Threat treatment earnings are also smaller in the LE condition, but not significantly so (22.97 and 25.47 ECU; p = 0.315).

4. CONCLUSION

Threats are common in human interaction and exchanges of threats often precede punishment. We have designed an experiment to study the effects of threats in a social dilemma setting in which the effect of punishment opportunities is well-understood, the Voluntary Contributions Mechanism. The Baseline treatment is a classical VCM game with sanctions. The Threat treatment includes a preliminary stage, in which participants can assign non-binding threats to punish, as a function of potential contribution levels of other agents. The Second Order treatment augments the Threat treatment with an opportunity to observe and to punish the differences between threats issued and actual punishment applied.

We find that threats are widely used. Most individuals threaten up to a high level of contribution. It appears that threats to punish high contributions are at least to some extent due to the fact that people use threats in an attempt to coordinate on a certain level of contribution, and not only to signal their willingness to punish behavior of which they

disapprove. While sanctions are much less severe than those that are threatened, the threats are nevertheless correlated with subsequent sanctions.

Our data provide no evidence that threats crowd out the intrinsic motivation to cooperate. In contrast, threats increase the average contribution level significantly. It appears that to some extent, threats are believed, and cooperation can be increased without the punishment being carried out. The ability to issue threats is welfare improving in the Threat treatment, but only in the latter periods of the session. This modest improvement in welfare results from the fact that the benefit to welfare of higher contributions is partly offset, at least in the early periods, by the cost of greater punishment.

When a discrepancy between threats and sanctions is observable to individuals, the effectiveness of threats vanishes completely. Here, individuals sanction those who fail to carry out their threats, and players moderate their threats as a result. The reduced level of threat, in turn, lowers contributions, and therefore overall welfare, returning then to levels even below those that would prevail in the absence of threats. The issuance of threats appears to induce a degree of expectation that they will be carried out. Failure to do so triggers a willingness-to-pay to punish the individual who issued the threat. The existence of this willingness to punish empty threats appears to be common knowledge. A failure to carry out threats, similarly to a failure to contribute or a failure to sanction non-cooperators, is treated as a punishable norm violation.

Since the beneficial effect of threats on welfare develops only in the latter periods of interaction, the results suggest that threats might not be effective in short-run relationships. We would need to observe a longer sequence of interaction (as in Gächter et al., 2008) to check whether the use of threats may be better suited to long-term interactions. In such a relationship, however, if threats and punishment can be associated, threats will only be effective if those who threaten are willing to follow through sufficiently often.

At least two types of extensions would refine our results and help establish their limitations. One would be to consider a greater level of punishment effectiveness, the

ratio of the cost of punishment to the sanctioning and the sanctioned parties, such as 1:3 or 1:4, instead of the 1:2 or 1:1 ratios that we specified here. Nikoforakis and Normann (2008) have shown that the availability of punishment promotes cooperation more strongly at greater punishment effectiveness ratios. We suspect that increasing the ratio would increase contributions to high levels in a Baseline treatment where threats are not possible. Thus, the use of threats would be more modest if they were available. Threats would be more likely to be carried out since they are cheaper, smaller, and it is less costly to punish discrepancies between threats and punishment assignments. Because contributions would be high and punishment low in all treatments, earnings would be high and more similar across treatments than for lower ratios. Thus, effective punishment technology may make threats unnecessary.

Another line of research could consider alternative implementations of threats. The method we used has the advantages that we can observe the relationship between threats and contribution levels, and that can be easily aggregated and summarized for recipients of the threats. A simpler method might be to have each individual set a threshold level of recipient contribution, below which they would administer a punishment of fixed magnitude. This provides a relatively clear message space and thus has the advantage of greater simplicity. However, it restricts punishment behavior a priori, ruling out graded threats and perverse punishment, so that important information about intended punishment is not available. The threshold itself might also create a focal point for contributions.

Alternatively, one might consider systems that have a richer message space. One possibility would be to allow players to condition punishment on deviations from average contributions, instead of absolute contribution levels. However, there is uncertainty about the average at the time that threats are issued, and individuals may also want to condition on absolute contribution levels. To permit conditioning on both would increase complexity substantially. Another possibility would be to allow unrestricted communication. This allows for positive and negative messages and greater freedom to express the intensity of a threat. However, there is less control, some threats may be vague or ambiguous, and it would not typically be possible to aggregate threats from

multiple individuals. An iterative process with a restricted message space, in which threats can be issued and updated in response to other threats, might be interesting, but is substantially more complex than the system we implemented, and the threats made in response to others' threats may be difficult to disentangle from those in response to contribution decisions.

REFERENCES

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

Andreoni, J. 1988. "Why Free Ride: Strategies and Learning in Public Goods Experiments," *Journal of Public Economics*, 35 (1), 57-73.

Bochet, O., T. Page, and L. Putterman. 2006. "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior and Organization*, 60 (1), 11-26.

Bochet, O., and L. Putterman. 2009. "Not just babble: Opening the black box of communication in a voluntary contribution experiment,"*European Economic Review*, 53 (3), 309-326.

Brosig, J., A. Ockenfels,and J. Weimann, 2003. "The effect of communication media on cooperation," *German Economic Review*, 4, 217–242.

Carpenter, J.P. 2007a. "Punishing Free Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods," *Games and Economic Behavior*, 60 (1), 31-51.

_____ 2007b."The demand for punishment,"*Journal of Economic Behavior and Organization,* 62, 522–542.

Cinyabuguma, M., T. Page, and L. Putterman. 2006. "Can Second Order Punishment Deter Perverse Punishment?,"*Experimental Economics*, 9 (3), 265-279.

Denant-Boemont, L., D. Masclet and C. Noussair. 2007. "Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment," *Economic Theory*, 31 (1), 145-167.

Dickinson, D., and M.C. Villeval. 2008. "Does Monitoring Decrease Work Effort? The Complementarity Between Agency and Crowding-Out Theories," *Games and Economic Behavior*, 63 (1), 56-76.

Duffy, J., and N. Feltovich. 2006. "Words, deeds, and lies: strategic behaviour in games with multiple signals ," *The Review of Economic Studies* 73 (3), 669-688.

Egas, M. and A. Riedl. 2008. "The economics of altruistic punishment and the maintenance of cooperation,"*Proceedings of the Royal Society B - Biological Sciences*, 275 (1637), 871-878.

Fehr E., and S. Gächter 2000. "Cooperation and Punishment in Public Goods Experiments,"*American Economic Review*, 90 (4), 980-94.

Fehr, E., and B. Rockenbach. 2003. "Detrimental Effects of Sanctions on Human Altruism," *Nature*, 422, 137-40.

Fischbacher, U. 2007. "Z-Tree: Zurich Toolbox for Ready-made Economic experiments,"*Experimental Economics*, 10 (2), 171-178.

Gächter, S., E. Renner, and M. Sefton. 2008. "The Long-Run Benefits of Punishment,"*Science*, 322, 5 December, 1510.

Herrmann B., Gächter S. and Thoeni C. (2008) "Anti-social punishment across societies", *Science* 319, 1362-1367 .

Houser, D.,E. Xiao, K. McCabe, and V. Smith. 2007. "Money, religion and revolution," *Economics of Governance*, 8 (1), 1-16.

_____. 2008."When Punishment Fails: Research on Sanctions, Intentions and Non-Cooperation,"*Games and Economic Behavior,* 62 (2), 509-532.

Isaac, R. M., K. McCue, and C. Plott. 1985. "Public Goods Provision in an Experimental Environment," *Journal of Public Economics*, 26 (1), 51–74.

Isaac, R. M., and J.M. Walker. 1988a. "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism," *Quarterly Journal of Economics,* 103 (1), 179-99.

_____. 1988b. "Communication and Free-Riding Behavior: The Voluntary Contributions Mechanism," *Economic Inquiry*, 26 (4), 585-608.

_____. 1991. "Costly Communication: An Experiment in a Nested Public Goods Problem," in T. Palfrey (Ed.). *Contemporary Laboratory Research in Political Economy*. Ann Arbor: Univ. of Michigan Press.

Kerr, N.L., and C.M. Kaufman-Gilliland. 1994. "Communication, commitment, and cooperation in social dilemmas," *Journal of Personality and Social Psychology*, 66, 513-529.

Krishnamurthy, S. 2001. "Communication Effects In Public Good Games With And Without Provision Points," in M. Isaac (Ed.). *Research In Experimental Economics*, Volume Eight, Amsterdam : JAI.

Ledyard J. 1995. "Public Goods: A Survey of Experimental Research", in Kagel J. and Roth. A., Eds., *Handbook of Experimental Economics*. Princeton: Princeton University Press, 111-194.

Li, J., E. Xiao, D. Houser, and P.R. Montague. 2009. "Neural responses to sanction threats in two-party economic exchange," *PNAS*, 106(39), 29 September, 16835-16840.

Marwell, G., and R.E. Ames. 1979. "Experiments on the provision of public goods. I: Resources, interest, group size, and the free-rider problem," *American Journal of Sociology,* 84 (6), 1335–1360.

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

Nikiforakis, N. 2008."Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?," *Journal of Public Economics*, 92, 91–112.

Nikiforakis, N. and H. Normann. 2008. “A Comparative Statics Analysis of Punishment in Public Goods Experiments”, *Experimental Economics,* 11, 358-369.

Noussair, C., and S. Tucker. 2005. “Combining Monetary and Social Sanctions to Promote Cooperation,” *Economic Inquiry*, 43 (3), 649-660.

Ones, U. and Putterman, L., 2007. "The ecology of collective action: A public goods and sanctions experiment with controlled group formation," *Journal of Economic Behavior & Organization*, vol. 62(4), 495-521

Ostrom, E., J. Walker, and R. Gardner. 1992. “Covenants With and Without a Sword: Self-Governance Is Possible,”*American Political Science Review*, 86 (2), 404–417.

Page T., L. Putterman and B. Garcia, 2008. "Getting Punnishment Right: Do Costly Monitoring or Redustributive Punishment Help?," Working Papers 2008-1, Brown University, Department of Economics.

Sefton, M., R. Shupp, and J. Walker, 2007. “The Effect of Rewards and Sanctions in Provision of Public Goods,” *Economic Inquiry*, 45, 671-690.

Yamagishi, T. 1986. “The Provision of a Sanctioning System as a Public Good,” *Journal of Personality and Social Psychology,* 51 (1), 110-116.

Table 1. Characteristics of the experimental sessions

Session number	*Number of participants*	*Number of groups*	*Treatment*	*Effectiveness of Punishment*
1	12	3	Baseline	High
2	16	4	Baseline	High
3	20	5	Threat	High
4	8	2	Threat	High
5	12	3	SdOrder	High
6	12	3	SdOrder	High
7	12	3	SdOrder	High
8	12	3	Baseline	Low
9	12	3	Baseline	Low
10	12	3	Baseline	Low
11	12	3	Threat	Low
12	12	3	Threat	Low
13	12	3	Threat	Low
14	12	3	SdOrder	Low
15	12	3	SdOrder	Low
16	12	3	SdOrder	Low
Total	200	50		

Table 2. Determinants of Threat Assignment: High Effectiveness Condition (random-effects Tobit models)

Dependent variable	Number of threat points assigned by i to j, j≠i				
	All *c*	For *c*=0	For *c*=10	For *c*=15	For *c*=19
	(1)	(2)	(3)	(4)	(5)
Threat treatment	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*
Second Order treatment	-0.837*	-4.547***	-1.944*	-1.905*	-4.399***
	(0.509)	(1.791)	(1.111)	(1.021)	(1.326)
Period	0.130***	0.245***	0.244***	0.391***	0.569***
	(0.007)	(0.027)	(0.018)	(0.018)	(0.026)
Final period	-0.930***	-2.844***	-2.006***	-2.649***	-3.163***
	(0.195)	(0.722)	(0.475)	(0.459)	(0.630)
Constant	3.983***	12.399***	5.591***	1.185	-3.386***
	(0.391)	(1.404)	(0.856)	(0.790)	(1.035)
# Obs.	3840	3840	3840	3840	3840
Left-censored	546	630	735	1065	1815
Right-censored	30	2169	1440	1068	828
Log-likelihood	-8166.982	-4823.275	-6514.949	-6650.759	-5496.098
ρ	0.684	0.763	0.701	0.670	0.657

Notes: *** Significant at the 0.01 level; ** at the 0.05 level; * at the 0.1 level. The "Threat treatment" variable is omitted as it is the reference category. The "Second Order treatment" variable is a dummy that equals 1 if the subject plays the Second Order treatment, and 0 otherwise. The "Period" variable is a time trend. "Final Period" is a dummy that equals 1 if the current period is the last one, and 0 otherwise.

Table 3. Determinants of the number of punishment points assigned by player *i* to player *j* in the two rounds of punishment: High Effectiveness condition (random-effects Tobit estimates)

Treatments	First round of punishment			Second round of punishment Second Order treatment			
		All treatments	All except Baseline				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Baseline treatment	*Ref.*	*Ref.*		-	-	-	-
Threat treatment	0.378	1.212*	*Ref.*	-	-	-	-
	(0.702)	(0.670)					
Second Order treat.	0.488	0.383	-0.893	-	-	-	-
	(0.657)	(0.624)	(0.582)				
Average contribution	-	-0.221***	-0.160***	0.095*	0.106*	0.114**	0.118**
of other group members (c_{-i})		(0.033)	(0.044)	(0.055)	(0.056)	(0.055)	(0.056)
Pos. diff. from group	-	-0.297***	-0.236***	0.039	0.030	-0.032	-0.046
average contrib.		(0.051)	(0.063)	(0.069)	(0.069)	(0.069)	(0.071)
Absolute neg. diff.	-	0.525***	0.407***	0.284***	0.288***	0.282***	0.281***
from group average contrib.		(0.021)	(0.027)	(0.028)	(0.028)	(0.028)	(0.028)
Threat *i* assigned to *j*	-	-	0.377***	-	-	-	
			(0.042)				
i is Anti-social threatener	-	-	1.364***	-	-	0.942*	0.949*
			(0.380)			(0.488)	(0.488)
j's average threat	-	-	-	-0.339	-0.573**	-0.576**	-0.573**
				(0.208)	(0.081)	(0.236)	(0.235)
j's average punishment	-	-	-	0.613***	1.278***	1.242***	1.240***
in first round				(0.125)	(0.354)	(0.357)	(0.358)
How much more *j* threatens	-	-	-	0.441**	0.681***	0.675***	0.673***
than he punishes				(0.218)	(0.246)	(0.247)	(0.246)
Sanctions *i* received	-	-	-	1.324***	1.651***	1.486***	1.559***
in first round				(0.345)	(0.379)	(0.381)	(0.382)
Sanctions *i* received					-0.504**	-0.536**	-0.519**
in first round**j*'s av. punishment					(0.231)	(0.231)	(0.230)
j is Anti-social punisher							0.604
							(0.558)
j is perverse punisher						0.619	
						(0.507)	
Pos. diff. between *j's* and	-	-	-	-	-0.182	-0.156	-0.155
average first round punishment					(0.261)	(0.261)	(0.261)
Abs. Neg. difference between	-				0.038	0.061	0.053
j's and average first round		-	-	-	(0.125)	(0.126)	(0.125)
Punishment							
Period	-	-0.314***	-0.329***	-0.177***	-0.169***	-0.154***	-0.155***
		(0.019)	(0.023)	(0.029)	(0.029)	(0.030)	(0.030)
Final period dummy	-	0.100	-0.455	-0.238	-0.218	-0.214	-0.225
		(0.567)	(0.695)	(0.938)	(0.936)	(0.923)	(0.924)
Constant	-6.902***	-0.051***	-0.224	-6.446***	-6.858***	-7.080***	-7.139***
	(0.549)	(0.728)	(0.923)	(1.104)	(1.146)	(1.135)	(1.139)
# observations	5520	5520	3840	2160	2160	2160	2160
# left cens.obs.	4676	4676	565	1892	1892	1892	1892
# right cens.obs.	46	46	30	-	-	-	-
						-	-
Log-likelihood	-3802.420	- 3212.445	-2204.581	-1069.524	-1066.825	1064.018	1064.167

Notes :*** Significant at the 0.01 level; ** at the 0.05 level; * at the 0.1 level. Standard errors are in parentheses. The "Baseline treatment" variable is omitted as it is the reference category. The "Threat (Second Order, respectively) treatment" variable is a dummy that takes 1 if the subject plays the Threat (Second Order, respectively) treatment, and 0 otherwise. The "Absolute neg. diff. from group average contrib." variable takes the absolute value

of the actual negative deviation of the subject's contribution from the others' average in case his own contribution is below the average. It takes zero otherwise. The variable "positive diff from group average contrib." is constructed analogously. The "Threat *i* assigned to *j*" variable is the number of threat points assigned by *i* to *j* for player *j*'s actual contribution. The "*j*'s average threat" variable indicates the average number of threat points assigned by *j* to the players other than *i*. The "*j*'s average punishment in first round" variable captures the average number of punishment points assigned by *j* to the players other than *i*. The "How much more j threatens than he punihes" takes the value of the difference between threats and punishment assigned by *j* in case *j* has assigned more threat points than punishment points, and 0 otherwise. The "Sanctions *i* received in first round" variable is dummy variable that takes 1 if *i* has received punishment points by the other players and 0 otherwise. The "*j* is anti-social punisher" variable takes value 1 if *j* assigned punishment points to those who contributed more than him, and 0 otherwise. The "*j* is perverse punisher" variable takes value 1 if *j* assigned punishment points to those who contributed more than the average of others, and 0 otherwise. The "Abs. neg. difference between *j's* and average first round punishment in first round" variable takes the absolute value of the actual negative deviation of the subject's punishment from the others' average in case *j* has punished less than the other players, and 0 otherwise. The variable "pos. diff between *j*'s and average first round punishment." is constructed analogously. The "Period" variable is a time trend. The "Final period" variable is equal to 1 if the observation corresponds to the final period of the game, and 0 otherwise.

Table 4. Determinants of contributions in the High Effectiveness condition

Models	*RE GLS*[a]	*RE GLS*[a]	*RE Tobit*[b]	*RE Tobit*[b]	*RE Tobit*[b]	*RE Tobit*[b]	*Tobit*
Treatments	*All*	*All except Baseline*	*All*	*All except Baseline*	*All except Baseline*	*All except Baseline*	*All Period 1*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Baseline	*Ref.*	-	*Ref.*	-	-	-	*Ref.*
Threat treatment	2.141*** (0.817)	*Ref.*	8.639*** (2.643)	*Ref.*	*Ref.*	*Ref.*	5.651*** (1.975)
Second Order Treatment	-0.098 (1.027)	-1.908** (0.829)	0.275 (2.427)	-7.673*** (2.519)	-8.095*** (2.382)	-8.123*** (2.387)	2.742 (1.975)
Average threat received	-	0.109*** (0.036)	-	0.306*** (0.079)	0.294*** (0.082)	0.268*** (0.080)	-
Threat received for *c*=20	-	-	-	-	-0.255** (0.115)		-
Ratio (threat rcvd. for c=20/threat rcvd. for c<20)						-5.812** (2.803)	
Threshold of threats assigned	-	-	-	-	0.191** (0.077)	0.197** (0.076)	-
Threat assigned for *c*=20	-	-	-	-	-4.851*** (1.585)	-4.894*** (1.583)	-
Period	0.055	-0.030	0.353***	0.308***	0.290***	0.292***	-
	(0.036)	(0.039)	(0.054)	(0.076)	(0.075)	(0.075)	-
Final period	-3.567***	-3.363***	-9.952***	-10.130***	-9.947***	-9.969***	-
	(0.750)	(0.953)	(1.389)	(1.742)	(1.732)	(1.731)	-
Constant	15.665***	16.158***	17.948***	22.121***	20.459***	20.704***	12.893***
	(0.595)	(0.645)	(1.890)	(2.241)	(2.293)	(2.304)	(1.360)
Observations	1840	1280	1840	1280	1280	1280	92
ρ	0.392	0.389	0.478	0.476	0.447	0.448	
Lef censored obs.			124	82	82	82	1
Right censored obs.			1073	798	798	798	30
Log likelihood			-3032.871	-1910.201	-1901.063	-1901.400	-233.961
R^2	0.044	0.100					

Notes: [a] Random-effects Generalized Least Squares model with robust standard errors clustered at the individual level in parentheses; [b] random-effects tobit; *** significant at the 0.01 level; ** at the 0.05 level; * at the 0.1 level. The "Baseline treatment" variable is omitted as it is the reference category. The "Threat (Second Order, respectively) treatment" variable is a dummy that takes 1 if the subject plays the Threat (Second Order, respectively) treatment, and 0 otherwise. The "Average threat received" variable is the sum of threat points received by a subject from his three group members. The "Threat received for *c*=20" variable is the sum of threat points received by a subject from his three group members for a potential contribution equal to 20. The "Threshold of threats assigned" is the contribution (between 0 and 20) from which a subject stops threatening others. The "Threat assigned for *c*=20" variable indicates the number of threat points assigned by the subject for a contribution equal to 20. The "Ratio threat" corresponds to the ratio "threat received for c=20"/threat received for c<20. The "Period" variable is a time trend. The "Final period" variable is equal to 1 if the observation corresponds to the final period of the game, and 0 otherwise.

Table 5. Determinants of payoffs in the HE condition (random-effects GLS models)

Dependent variables	Before-sanction payoffs	After-sanction payoffs	After-sanction payoffs	After-sanction payoffs Periods 11-20	After-sanction payoffs Periods 15-20
	(1)	(2)	(3)	(4)	(5)
Baseline treatment	*Ref.*	*Ref.*	*Ref.*	*Ref.*	*Ref.*
Own contribution	-0.303***	0.334***	0.326***	0.212***	0.172***
	(0.040)	(0.065)	(0.067)	(0.064)	(0.059)
Threat treatment	1.935***	-1.079	-2.597	0.799	1.752*
	(0.666)	(1.568)	(2.198)	(1.122)	(1.048)
Threat treat.*last 10 periods			3.070**		
			(1.567)		
Second Order treatment	-0.089	-2.605*	-4.207*	-0.976	-0.866
	(0.884)	(1.589)	(2.277)	(1.122)	(1.159)
Second Order treatment *last 10 periods			3.201*		
			(1.647)		
Periods 11-20	-0.056	4.629**	2.442**		
	(0.314)	(0.735)	(0.894)		
Constant	34.487***	18.160***	19.383***	23.617***	24.376
	(0.924)	(1.788)	(1.949)	(1.483)	(1.357)
# of observations	1840	1840	1840	552	552
R^2	0.17	0.130	0.135	0.081	0.095

Note: *** significant at the 0.01 level; ** at the 0.05 level; * at the 0.1 level. Robust standard errors in parentheses with clustering at the individual level. The "Baseline treatment" variable is omitted as it is the reference category. The "Threat (Second Order, respectively) treatment" variable is a dummy that takes 1 if the subject plays the Threat (Second Order, respectively) treatment, and 0 otherwise. The "Period 11-20" variable is equal to 1 if the observation belongs to the last ten periods of the experiment, and 0 otherwise.

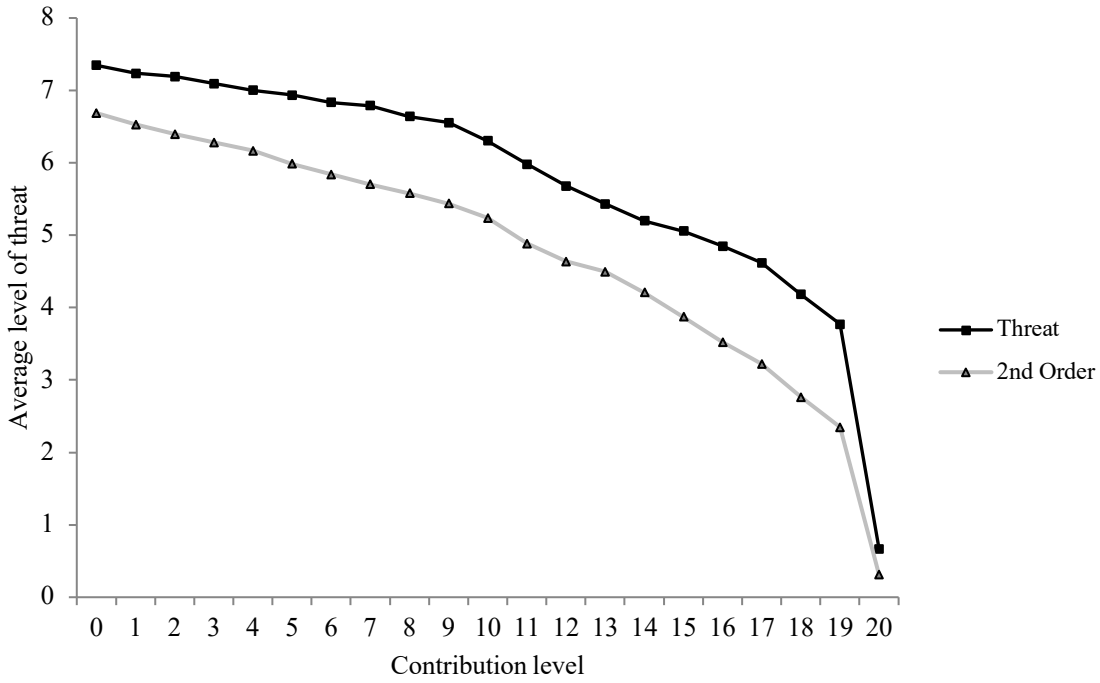


Figure 1. Average number of threat points assigned for each contribution level by treatment in the High Effectiveness condition

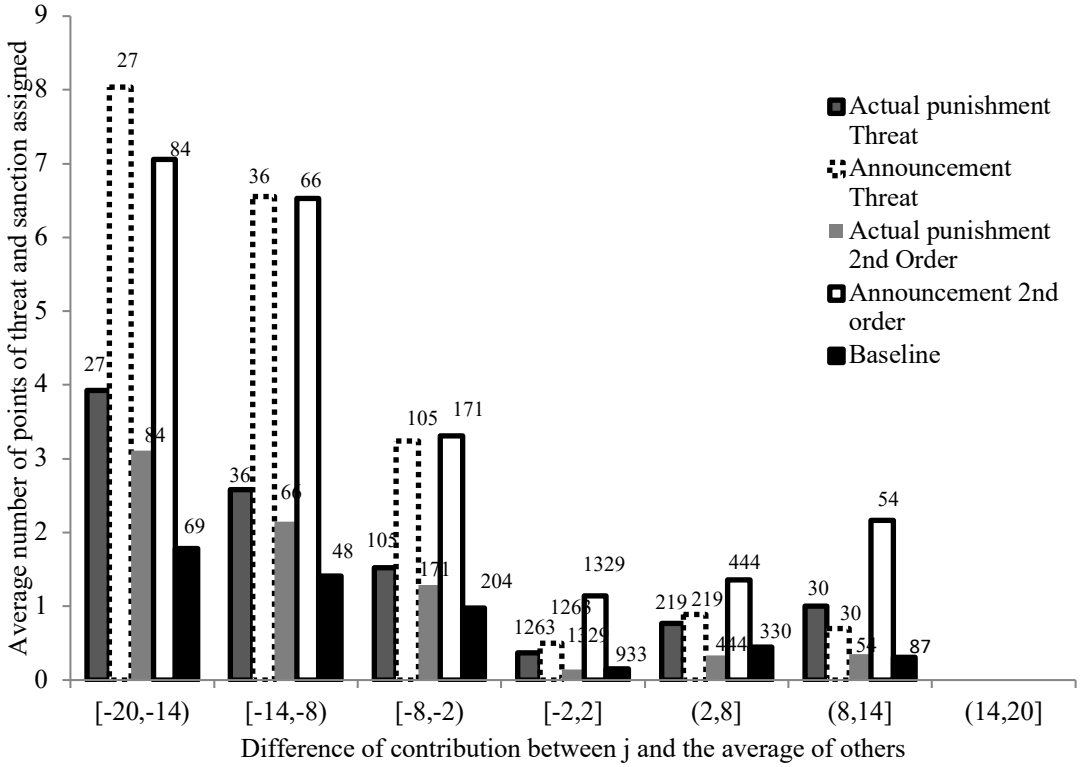


Figure 2. Average individual threat and actual punishment by treatment and by category of difference between recipient *j*'s and the average group contribution, in the High Effectiveness condition, all treatments

Figure 2 shows the average number of threat points issued for the actual subsequent realized contribution levels and the actual number of punishment points assigned, in the second stage of both the Threat and the Second Order treatments, as well as the average number of points assigned in the Baseline treatment. These are displayed as a function of the difference between the target's contribution and the average group contribution (excluding target *j*'s own contribution). Figure 2 should be interpreted while taking into account that contributions relative to the group average were not yet known at the time of threat assignment.

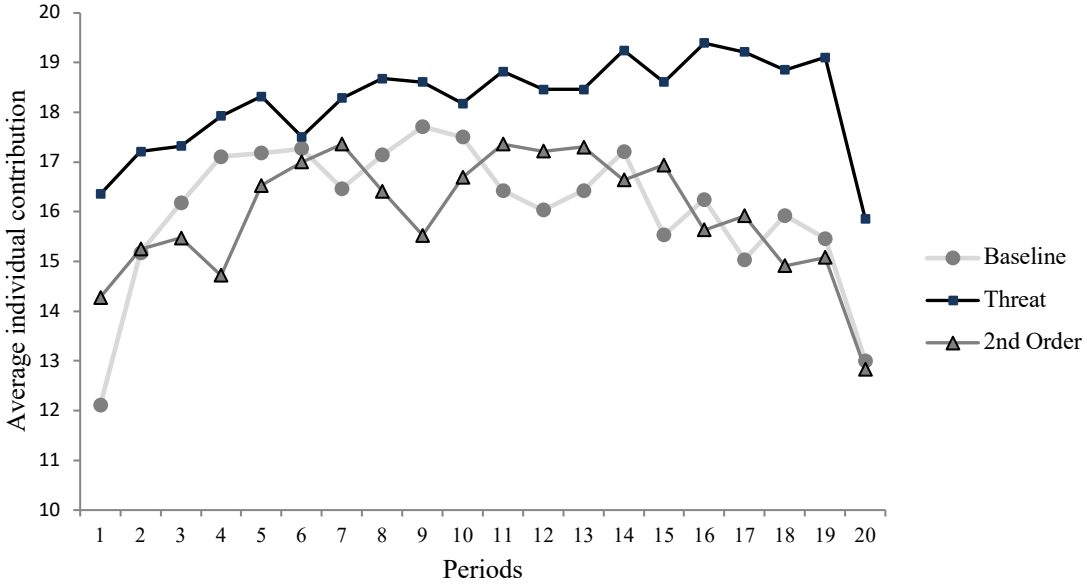


Figure 3. Average individual contributions over time by treatment in the High Effectiveness condition

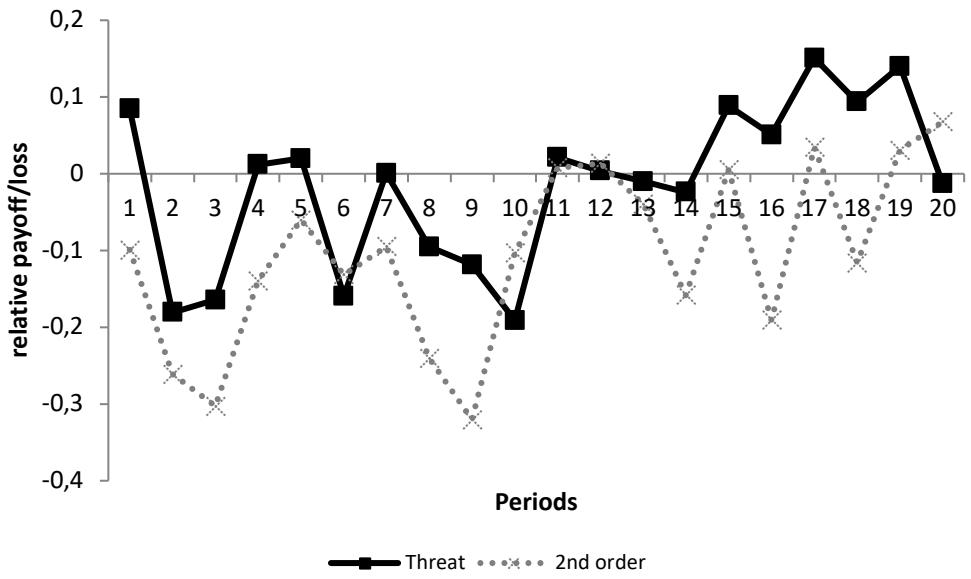


Figure 4. Average payoff difference between Threat and Second order treatments relative to the Baseline treatment: High Effectiveness condition

The payoff difference is calculated as *(treatment-Baseline)/Baseline*. For instance, in the Threat treatment, payoffs are roughly 15 percent greater than in the Baseline treatment in period 17. The payoff is about 30 percent lower in period 3 of the Second Order treatment compared to the same period of Baseline.

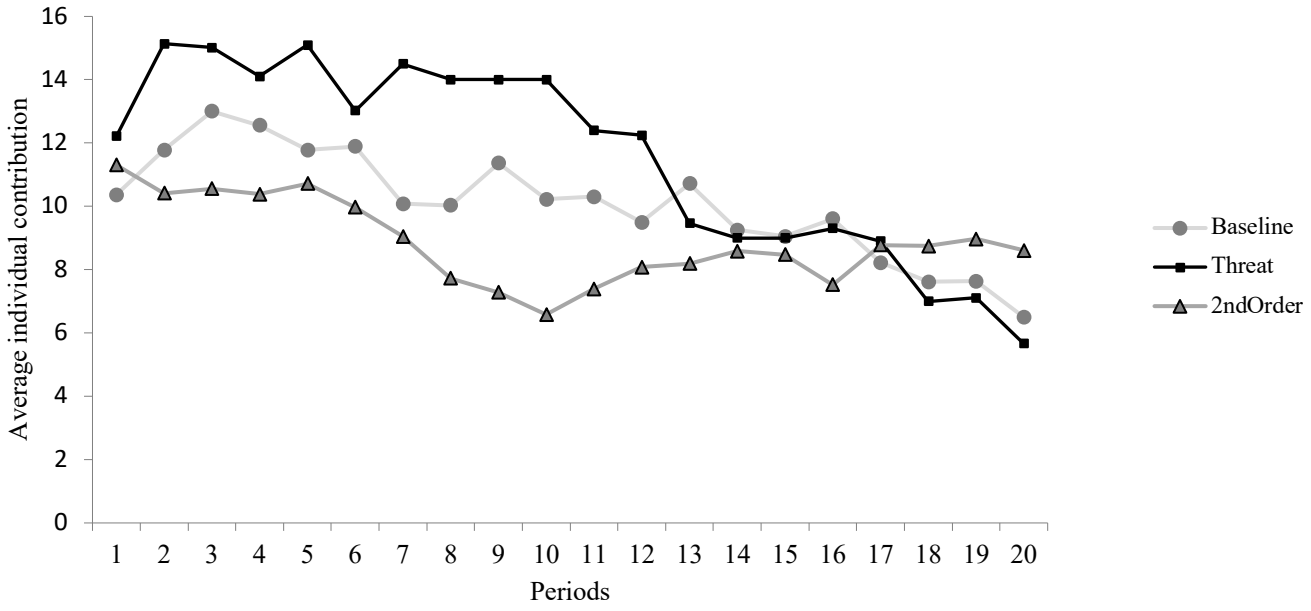


Figure 5. Average individual contributions over time by treatment
in the Low Effectiveness condition

Online appendix

Appendix. Instructions of the Threat treatment (high effectiveness condition)*(Translated from the original French text. The instructions for the other treatments are available upon request)*

You are taking part in an economic experiment, during which you can earn money. Your earnings depend on your decisions and on the decisions of the other participants with whom you will interact. It is therefore important to read these instructions with attention.

All of the transactions during the experiment and your entire earnings will be calculated in ECU (Experimental Currency Units). At the end of the experiment, the total amount of ECU you have earned during this session will be converted to Euros, and paid to you in cash in a separate room. You will be paid by somebody who is not aware of the content of the experiment, according to the following rules:

- Your final payoff in ECU consists of the total of your payoffs in each of the 20 periods that make up this session.
- This final payoff in ECU will be converted into Euros at the rate: 100 ECU = 2 Euros.
- In addition, you will be given a show up fee of 5 Euros.

At the beginning of the session, the participants are divided into groups of four. You will therefore interact with three other participants. **During the 20 periods, you will interact with the same persons**. You will never be informed of the identity of these persons.

Description of each period

In each period, after receiving an endowment of 20 ECU each, the four participants belonging to a group can participate in a project, by contributing to a group account that will be shared among them. The amount of this group account is determined by the total of the individual contributions of the four members of the group. Next, the group members can indicate their disapproval of the contribution of other members of the group by assigning points that reduce their payoff. Each period consists of three stages:

- During the first stage, each group member indicates how many disapproval points he would be ready to assign to other group members for each possible contribution level in the second stage.
- During the second stage, after being informed of the number of disapproval points that the other group members propose to assign for each possible contribution level, each of the four group members decides simultaneously on his actual contribution to the project.
- During the third stage, after being informed of the individual contributions of the other group members, each one decides on the number of disapproval points he actually assigns to other group members and their payoffs are reduced accordingly.

The details of each stage are described below.

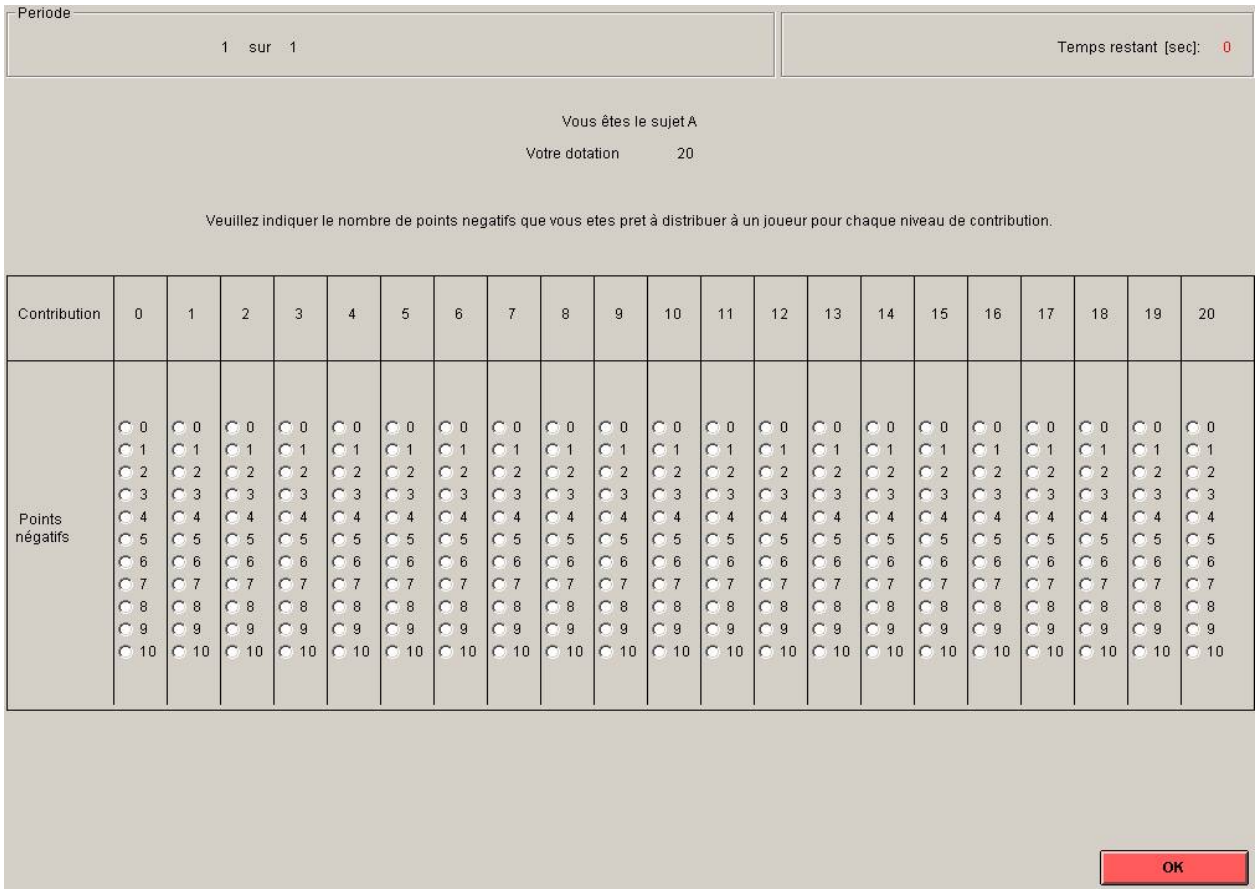
First stage

You announce the number of points you would like to assign to each other group member for each possible contribution level (between 0 and 20 ECU)to the project in the second stage. **The number of points you announce for a group member indicates your degree of disapproval for each contribution level (from 10 points for the highest disapproval to 0 point for no disapproval).** The three other members of your group are informed of your announcement before they decide on their contribution level.

For the moment, the points you announce affect neither your payoffs nor the payoffs of your group members. They simply indicate to the others your willingness to reduce their payoffs for each possible contribution amount. It is only after every group member will have decided his contribution during the second stage that you will, in the third stage, confirm or modify your announced number of points. These points will then affect both your payoffs and the payoff of your group members, as indicated below.

- You announce the number of points that you would be willing to assign for each possible contribution level of the other members of your group. You must enter a number, between 0 and 10, for each possible contribution. If you do not want express disapproval, you must enter 0.
- At the end of the first stage, the number of points you would be willing to assign for each contribution level will be announced to the members of your group. You are also informed of the total number of points that your three group members are willing to assign to you in the third stage for each of your possible contribution levels.

Below is the screenshot for the first stage.



Second stage

You receive an endowment of 20 ECU. After being informed of the total number of points that you are susceptible to receiving from the other group members for each possible contribution level, you decide on your contribution to the project.

You, as well as the three group members decide simultaneously, how much of your endowment you will allocate to the project, by indicating a number between 0 and 20. To validate your choice, click the OK button.

After all group members have made their decision, your screen will show you the total amount of ECU contributed to the project by the members of your group (including your contribution). You are also informed of your current payoff at this stage.

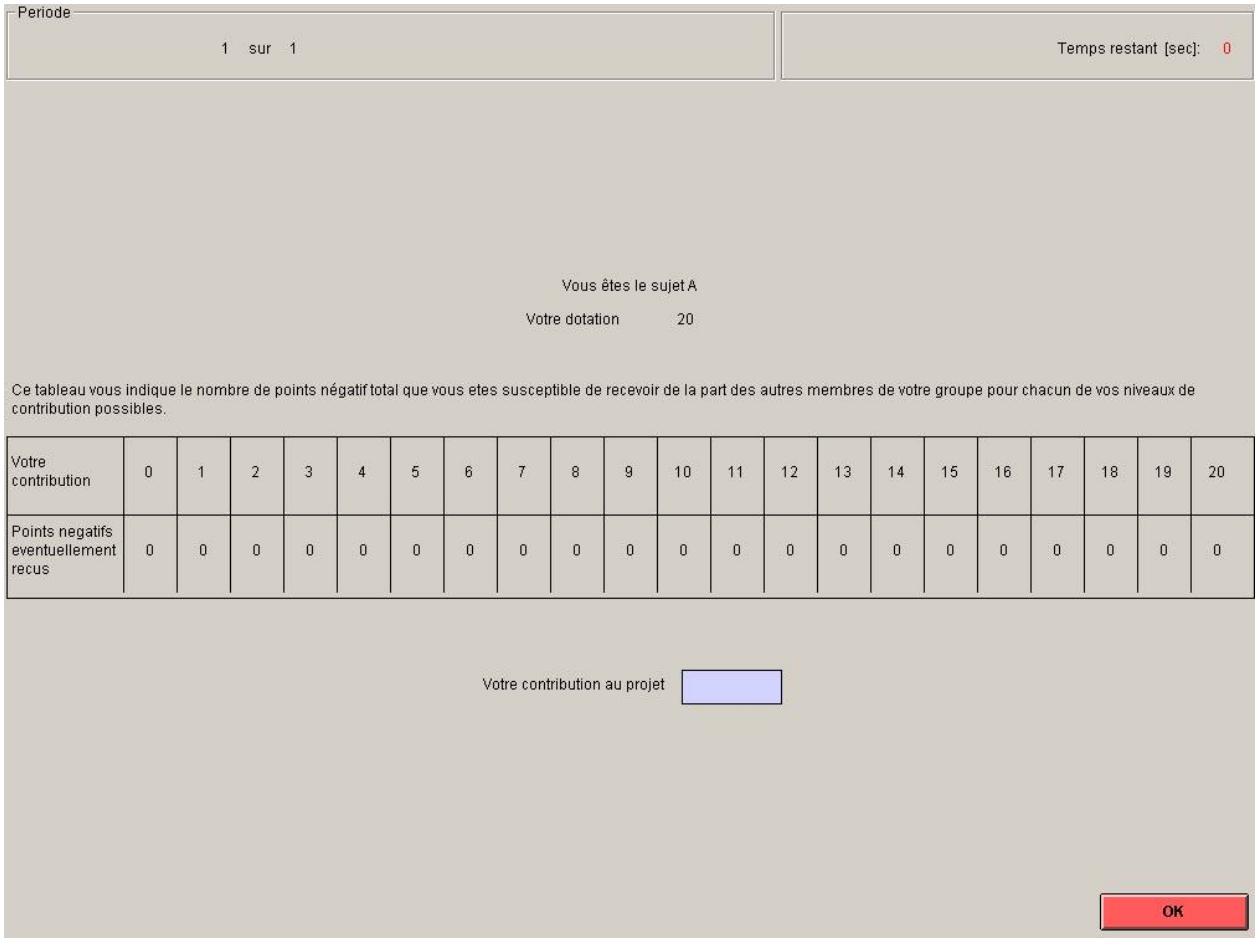
Your payoff at this second stage consists of two parts:

- the amount of your endowment which you have kept for yourself (that is, 20 – your contribution to the project),
- your income from the project: this income represents 40% of the total contribution of all four group members to the project .

Your payoff in ECU in this second stage is computed by the program as follows:

(20-your contribution to the project) + 40%*(total contributions of the group to the project)

Below is the screenshot for the second stage.



The payoff of each group member is calculated in the same way, which means that each group member receives the same income from the project.

For example, suppose that the total of the contributions of all group members is 60 ECU. In this example each member of the group receives a second-stage payoff from the project of 40% (of 60 ECU) = 24 ECU. On the other hand, if the total contribution to the project is 9 ECU, then each member of the group receives 40% (of 9 ECU) = 3.6 ECU from the project.

For each ECU of your endowment that you keep for yourself you earn an income of 1 ECU. Every ECU you contribute to the project instead increases the total contribution to the project by one ECU. The income from the project will increase by 0.4 ECU per person and so, the total income of the group from the project will rise by 1.6 ECU. This means that your contribution to the project also increases the income of the other group members.

On the other hand you will earn money from each ECU contributed by the other members to the project. For each ECU contributed by any group member you earn 40% (1) = 0.4 ECU.

Third stage

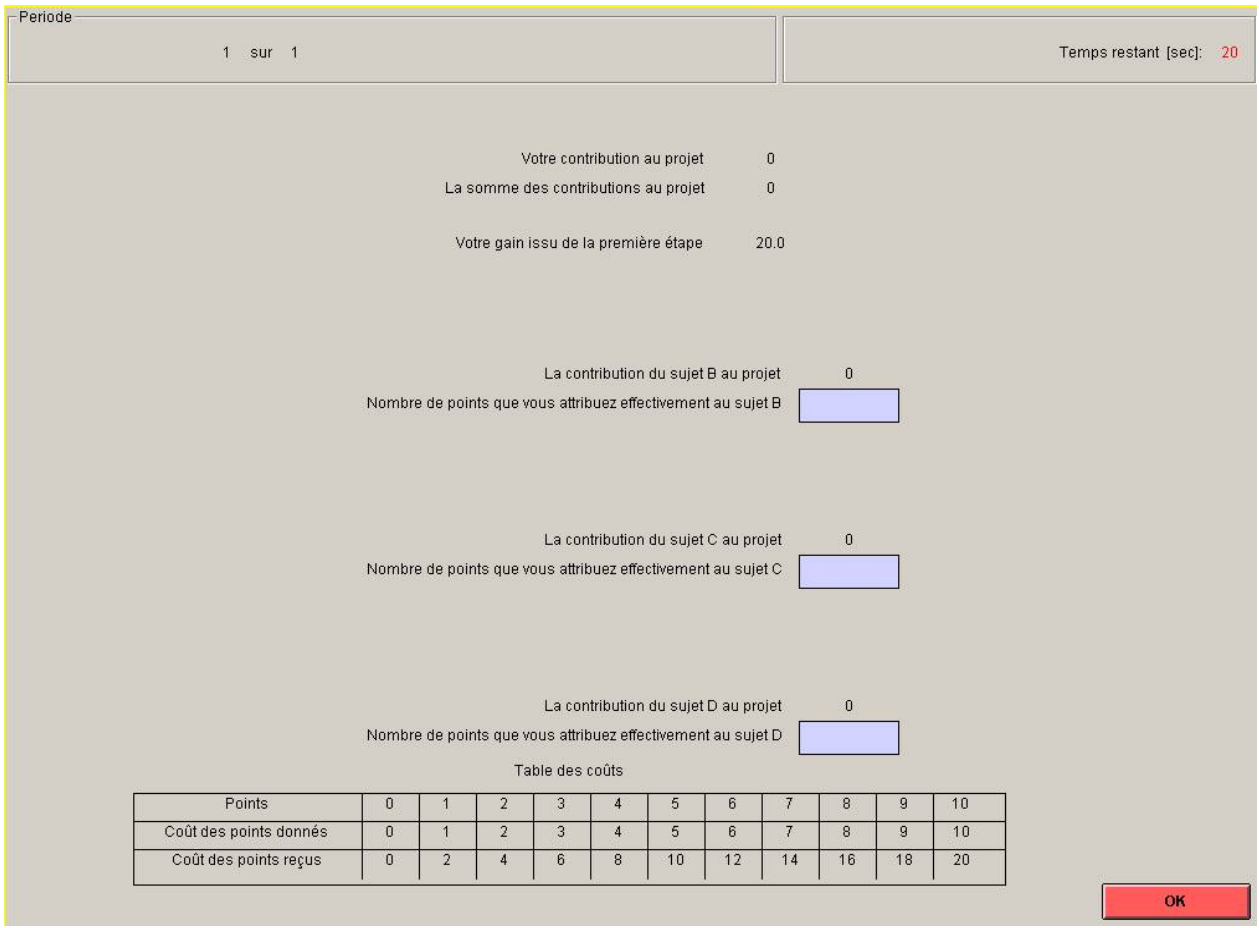
After being informed of the contribution of each of the other members of your group, you can, if you would like, reduce or leave unchanged their payoff by assigning points. **This number of points can be the same or different from the number you have announced in the first stage.** You can assign a particular number of points to a member of your group to express a level of disapproval (10 points for the highest disapproval, 0 points for no disapproval). Each point assigned to a particular group member reduces her second-stage income by two points.
Your decision during the third stage depends on the actual contributions and can change both your payoff and the payoff of the other members of your group. Similarly, your payoff can be changed if the other group members wish to do so.

- You are informed of the contribution of each of the other three members of your group to the project in the second stage of the game. Note: the order in which each contribution is displayed changes randomly in each period (in other words, for example, the number that appears first on your screen does not always correspond to the decision of the same player).

- You decide next on how many points to send to each of the other three members of your group to reduce their payoff or leave it unchanged. Each point assigned to a group member reduces his second-stage payoff by 2 ECU. If you assign 0 point to another member, you do not change hissecond-stage payoff. If you assign 1 point to a group member, you reduce his second-stage payoff by 2ECU; if you assign 2 points, you reduce his second-stage payoff by 4 ECU; etc. You must enter a value for each member, between 0 and 10 points. If you do not wish to reduce the payoff of a specific member, then you must enter 0.

- If you assign points, you pay a cost that depends on the number of points you assign to each subject. Each point you assign reduces your second-stage payoff by 1 ECU. Your total cost is equal to the sum of the costs of assigning points to each of the other three group members. If you assign two points to one group member, it will cost you 2 ECU. If you assign 9 points to another member, it will cost you 9 ECU more. If you give the last group member no points, it does not cost you anything. In this example, the total cost of the assigned points is 11 ECU (2+9+0). These costs will be displayed on your screen. You can modify your decisions until you click the OK button.

Below is the screenshot for the third stage.



Points	0	1	2	3	4	5	6	7	8	9	10
Coût des points donnés	0	1	2	3	4	5	6	7	8	9	10
Coût des points reçus	0	2	4	6	8	10	12	14	16	18	20

- Your final payoff in ECU in each period is calculated by the computer as follows:

Final payoff = (second stage payoff)-cost of points received in the third stage-cost of points assigned in the third stage

Note that in the calculation of payoffs, the cost of points received cannot exceed your second-stage income.

For example, if you received a total of 3 points from the other three members of your group, your second-stage payoff is reduced by 6 ECU. If you received 4 points,your second-stage payoff is reduced by 8 ECU. If you received 10 points, you lose 20 ECU of your second-stage payoff.

Your third-stage payoff can therefore be negative if the cost of the points you have assigned exceeds your second-stage payoff net of the cost of received points. You can, however, avoid such losses with certainty through your own decisions.

To summarize

Each period consists of three stages.

- In the first stage, you announce the number of points you would be ready to assign to your group members for each possible contribution level. The three group members are informed of your announcement. Similarly, you are informed of the total numbers of points announced by your three other group members for each possible contribution.

- In the second stage, you choose your contribution to the project.

- In the third stage, you are informed of the individual contribution of each member of your group. You can assign points that will reduce their payoff and that can differ or not from your announcement in stage 1.

At the end of each period, the next period starts automatically. You receive a new endowment of 20 ECU.

Thank you for answering the questionnaire that has been distributed; we will check your answers individually. If you have any questions about these instructions, please raise your hand. We will answer your questions in private.

Communicating with the other participants during the experiment is strictly forbidden at the risk of being excluded from the session and from receiving your payment.

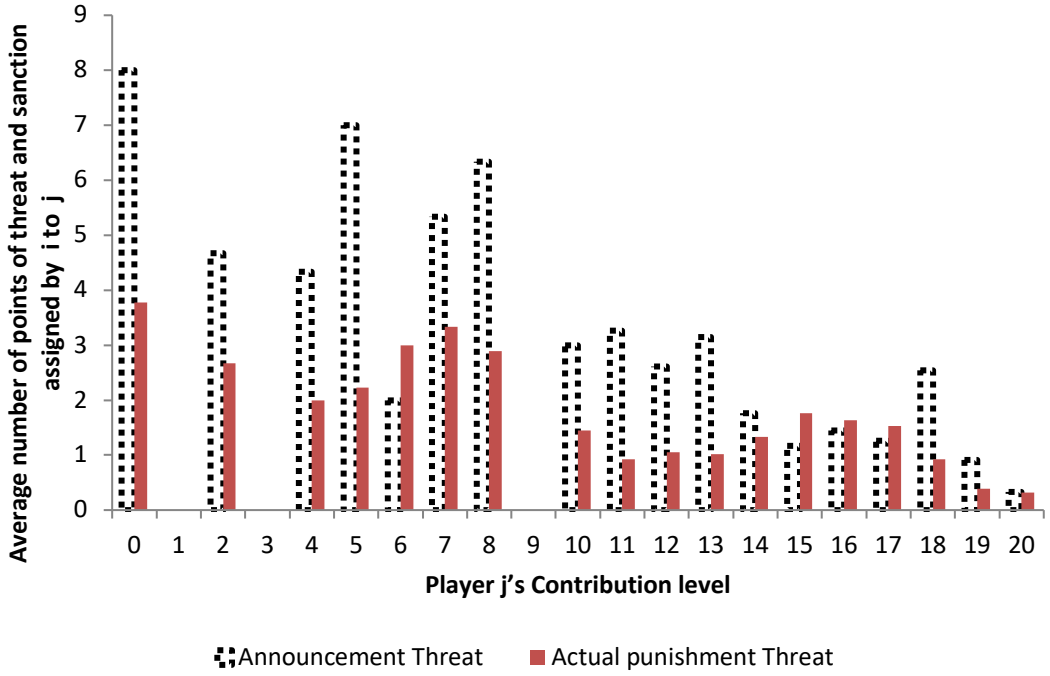


Figure A1. Average individual threat and actual punishment in the Threat treatment by absolute contribution level

Figure A1 displays the average number of threat points issued for the actual subsequent realized contribution levels, and the actual number of punishment points assigned, in the Threat treatment. These are displayed as a function of the target's contribution.

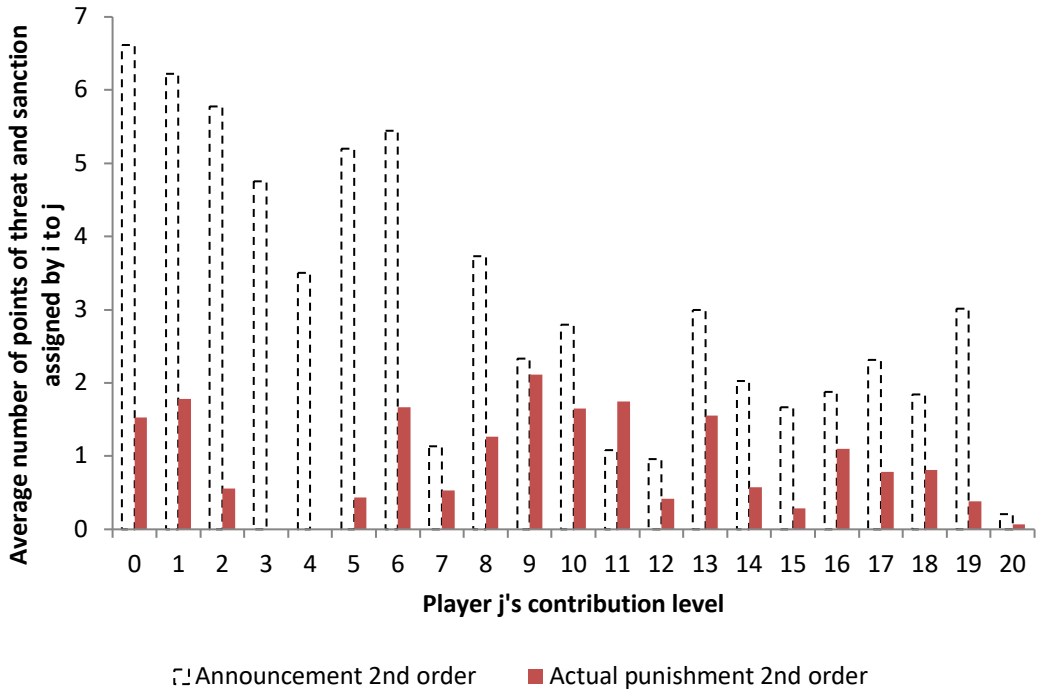


Figure A2. Average individual threat and actual punishment in the 2nd order treatment by absolute contribution level

Figure A2 displays the average number of threat points issued for the actual subsequent realized contribution levels and the actual number of punishment points assigned in the second stage of the second order treatment. These are displayed as a function of the target's contribution.

Table A1. Determinants of the contribution threshold above which the player no longer threatens to punish

Dependent variable: contribution threshold	
Model: Random-Effects Tobit Model	
Threat treatment	*Ref.*
Second Order treatment	-0.497
	(1.164)
Period	0.085***
	(0.018)
Final period	-2.112***
	(0.471)
Constant	14.797***
	(0.895)
# Obs.	3840
Left-censored	551
Right-censored	336
Log-likelihood	-10331.63
ρ	0.663

Notes: *** Significant at the 0.01 level; ** at the 0.05 level; * at the 0.1 level. The "Threat treatment" variable is omitted as it is the reference category. The "Second Order treatment" variable is a dummy that equals 1 if the subject participates in the Second Order treatment, and 0 otherwise. The "Period" variable is a time trend. "Final Period" is a dummy that equals 1 if the current period is the last one, and 0 otherwise.

Table A2 Determinants of the probability of being anti-social punisher (i.e. of punishing those who contributed more than the punisher).

.

Dependent variable: player *i* is anti-social punisher	
Model: Random-Effects Probit Model	
Threat treatment	*Ref.*
Second Order treatment	-0.497
	(1.164)
Player *i*'s Contribution	-0.073***
	(0.006)
i is anti-social threatener	0.602***
	(0.095)
Period	-0.096***
	(0.07)
Final period	0.143
	(0.226)
Constant	-0.957***
	(0.165)
# Obs.	8160
Log-likelihood	-1081.09
ρ	0.624

Notes: *** Significant at the 0.01 level; ** at the 0.05 level; * at the 0.1 level. The "Threat treatment" variable is omitted as it is the reference category. The "Second Order treatment" variable is a dummy that equals 1 if the subject participates in the Second Order treatment, and 0 otherwise. The "Period" variable is a time trend. "Final Period" is a dummy that equals 1 if the current period is the last one, and 0 otherwise. The "*i* is Anti-social threatener" variable takes value 1 if *i* assigned threat points for the highest possible contribution level, and 0 otherwise.

Table A3. Determinants of changes in the threats made between periods *t* and *t+1*

Dependent variable: change in threat between *t* and *t*+1 Model: GLS		
	Indiv. who threaten more than they punish in period t.	*Indiv. who threaten less than they punish in first-round punish. of period t.*
Sanctions *i* received in second round of punishment in period *t*	-0.170** (0.077)	0.033 (0.059)
Diff. between *i*'s threat and first-round sanction in period *t*	-0.452*** (0.088)	0.032 (0.138)
Constant	0.771** (0.303)	0.798*** (0.132)
Observations	711	1341
R^2	0.124	0.002

Note: * significant at 10%; ** significant at 5%; *** significant at 1%; Robust standard errors in parentheses; cluster (id)

Table A4. Determinants of changes in individual contributions between periods *t* and *t+1*

Dependent variable: change in contribution between *t* and *t*+1				
Treatments	All	2nd order	All	2nd order
Model: GLS				
	Indiv. who contribute less than the group average in period t		*Indiv. who contribute more than the group average in period* t	
Sanctions *i* received in first round of punishment in period *t*	0.316***	0.297***	-0.087	0.442**
	(0.096)	(0.112)	(0.184)	(0.190)
Sanctions *i* received in first round of punishment in period *t* *Threat treatment	-0.036		0.028	
	(0.122)		(0.188)	
Sanctions *i* received in first round of punishment in period *t**2nd. order treatment	0.073		0.177	
	(0.106)		(0.199)	
Sanctions *i* received in second round of punishment in period. *T*		-0.248		-0.319
		(0.185)		(0.264)
Diff. between *i*'s and average contribution sanction in period *t*	-0.543***	-0.690***	-0.447***	-0.505***
	(0.085)	(0.115)	(0.074)	(0.106)
Period	-0.129	-0.070	-0.092***	-0.065
	(0.046)	(0.084)	(0.029)	(0.055)
Constant	0.477**	0.280	0.509	0.142
	(0.498)	(0.808)	(0.434)	(0.738)
Observations	457	198	1291	486
R^2	0.428	0.546	0.081	0.101

Note : * significant at 10%; ** significant at 5%; *** significant at 1%;
Robust standard errors in parentheses; cluster (id)