

IZA DP No. 5689

## **New Insights into Conditional Cooperation and Punishment from a Strategy Method Experiment**

Stephen L. Cheung

May 2011

# **New Insights into Conditional Cooperation and Punishment from a Strategy Method Experiment**

**Stephen L. Cheung**

*University of Sydney  
and IZA*

Discussion Paper No. 5689

May 2011

IZA

P.O. Box 7240  
53072 Bonn  
Germany

Phone: +49-228-3894-0

Fax: +49-228-3894-180

E-mail: [iza@iza.org](mailto:iza@iza.org)

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

## ABSTRACT

### **New Insights into Conditional Cooperation and Punishment from a Strategy Method Experiment\***

This paper introduces new experimental designs to enrich understanding of conditional cooperation and punishment in public good games. The key to these methods is to elicit complete contribution or punishment profiles using the strategy method. It is found that the selfish bias in conditional cooperation is made significantly worse when other players contribute more unequally. Contingent punishment strategies are found to increase with decreasing contributions by the target player and also increasing contributions by a third player. “Antisocial” punishments are not directed specifically toward high contributors, but may be motivated by pre-emptive retaliation against punishment a player expects to incur.

JEL Classification: C72, C91, D70, H41

Keywords: conditional cooperation, selfish bias, punishment, public good experiment, strategy method

Corresponding author:

Stephen L. Cheung  
School of Economics  
University of Sydney  
Merewether Building H04  
Sydney NSW 2006  
Australia  
E-mail: [Stephen.Cheung@sydney.edu.au](mailto:Stephen.Cheung@sydney.edu.au)

---

\* I thank Michele Bernasconi, Bram Cadsby, Gary Charness, Ananish Chaudhuri, Simon Gächter, Glenn Harrison, Danielle Merrett, Nikos Nikiforakis, Charles Noussair, Stefan Palan, Robert Slonim, Marie Claire Villeval, Tom Wilkening, seminar audiences at The University of Sydney and Tilburg University, and participants at the VII LabSi Workshop on Experimental and Behavioural Economics in Siena in April 2010, the International Conference of the Economic Science Association in Copenhagen in July 2010, and the Fifth Australia and New Zealand Workshop on Experimental Economics in Sydney in November 2010. I also thank Tim Capon and Min-Taec Kim for lab assistance, and The University of Sydney for financial support. A previous draft of this paper circulated under the title “Conditional punishment and cooperation: A strategy method public good experiment”.

## INTRODUCTION

Understanding the nature of the preferences that sustain cooperative behaviour, when individual material payoffs can be improved by defecting, is a fundamental question in many fields of economics, and indeed the social sciences more generally. The model of voluntary contribution to a public good provides a simple metaphor for many social dilemmas in which contribution is socially efficient, but where agents motivated by material self-interest have unilateral incentives to free-ride. In purely one-shot interactions – with no prospect for reputation, retaliation or other longer-term strategic considerations – standard theory predicts universal free-riding as the dominant strategy equilibrium. Yet a large body of empirical evidence from controlled experiments suggests a more complex picture.<sup>1</sup> This research finds that while many people do indeed free-ride, there are others who contribute a not-inconsequential share of their resources toward public goods, even in one-shot interactions.

A key insight from this literature is that many of those who are willing to contribute are in fact conditional cooperators, who prefer to contribute only when others do likewise. A seminal paper in this regard is by Fischbacher, Gächter and Fehr (2001), hereinafter FGF.<sup>2</sup> Their experiment elicits conditional contribution decisions in response to a range of scenarios regarding the average contributions of others, and finds a relationship that is positive but less than one-for-one. In particular, even subjects who are classified as conditional cooperators display a “selfish bias” in their contribution behaviour. One consequence of selfish bias is that cooperation is fragile and prone to dissipate even in repeated interactions (Fischbacher and Gächter, 2010). Because of this fragility, there is considerable interest in institutions that can strengthen cooperation in the face of the temptation to free-ride.

One such institution that has attracted considerable interest is the opportunity for players to discipline one another through informal peer sanctions. Fehr and Gächter (2000, 2002) model this experimentally by providing feedback to subjects on the contributions of their counterparts and then allowing them, at a cost to themselves, to assign targeted reductions in the earnings of others. If those who contribute are also willing to use punishment to discipline free-riders, then cooperation might be stabilised. Once again however, in a one-shot interaction, standard theory predicts that because there is no future benefit that can accrue

---

<sup>1</sup> See Chaudhuri (2011), Gächter and Herrmann (2009), and Ledyard (1995) for excellent reviews of this literature.

<sup>2</sup> See also Keser and van Winden (2000) as well as the additional references cited in footnote 12 below, and the survey by Chaudhuri (2011).

from imposing costly punishment upon others, no player motivated by material self-interest should ever choose to do so. That being the case, the opportunity to punish should also have no effect upon contributions. To the contrary, Fehr and Gächter find that many subjects are indeed willing to punish, and that contributions are higher as a result, even when there is little or no prospect of interacting with the same counterparts again. Given that both conditional cooperation and punishment are at odds with conventional theory, especially in one-shot settings, they have helped to stimulate a lively literature on models of social preferences.<sup>3</sup>

This aim of this paper is to redress significant limitations of existing studies both without and with punishment, and to introduce new designs that enrich understanding of how players' willingness to either conditionally cooperate or punish vary with the cooperativeness of others. The key to these methods is the systematic application of the "strategy method" of elicitation (Selten, 1967). In a strategy method design, each subject specifies a complete profile of choices (either contributions or punishments, as the case may be) in response to every possible combination of the choices of others. As will be explained shortly, previous studies of games without punishment – in particular FGF and the studies that replicate and build upon it – only apply a limited form of the strategy method. As a result, they overlook important nuances in how conditional cooperation responds to inequality in the contributions of others. Likewise, previous studies of punishment in public good experiments do not make use of the strategy method at all.

In the context of a public good experiment, the inherent difficulty of the strategy method arises from the very large number of potential combinations of contributions. For the standard parameters used by Fehr and Gächter (2000, 2002), there are 21 possible integer levels of contribution from 0 to 20. Since each subject is assigned to a group of four, there are  $21^3 = 9,261$  possible combinations of the contributions of the other three players. In a game with punishment, each subject would then have to decide how much punishment to assign to each of the three other players, so there would be  $9,261 \times 3 = 27,783$  different punishments that need to be specified! Clearly, to make the strategy method operational, it is necessary to simplify the strategy space of the game.

To apply a form of the strategy method to a game *without* punishment, FGF reduce the number of conditional contribution decisions that need to be elicited from 9,261 to 21 by asking each subject to specify a vector of contributions conditional upon each possible value

---

<sup>3</sup> See Cooper and Kagel (in press) for a survey of the interplay between theory and experiments in this area.

of the *average* contribution of the other players, rounded to the nearest integer. Prior to doing this, each subject also specifies an unconditional contribution decision. Afterwards, a random draw selects one player from each group whose contribution will be determined by the contribution table, whereas for the others the unconditional contribution is binding. Using this procedure, it is possible to elicit conditional contribution decisions from every subject in an incentive compatible manner.

FGF document selfish bias even among the subjects they classify as conditional cooperators, since these subjects tend to fall short of matching the average contributions of others. However, because FGF only elicit conditional contributions as a function of the average of others' contributions, their design does not enable them to detect how conditional contributions might vary in response to changes in the *composition of that average*. In short, *they do not truly elicit contribution strategies* in response to every possible combination of the other players' actions. For a given average of the other players' contributions, it is quite conceivable that conditional contributions could be quite different in response to a scenario in which all contribute equally, compared to one in which the same average is accounted for entirely by the contribution of a single individual. To address questions of this kind using the strategy method, it is necessary to elicit conditional contributions in response to *combinations of contributions* of the other group members, and not only to averages.

Previous studies of punishment in public good experiments rely on the traditional "direct-response" method of elicitation: subjects are simply asked to make punishment decisions in response to the *actual specific contributions* of the other group members with whom they are matched. The problem is that this information is only ever sought, and thus revealed, for those patterns of contribution that in fact arise in the course of play of the game. That is, the observable variation in contribution behaviour is limited to the actual choices made by others. As a result, it is impossible to determine how a punisher's behaviour might differ in the face of some alternative counterfactual pattern of contributions.

More data, and possibly greater variation, can be obtained by pooling decisions from multiple rounds of a repeated game. However, depending upon the matching protocol, repeated interaction can bring additional strategic considerations into play and, in any case, it need not ensure that all subjects are exposed to a full range of variation in the contributions of others. In short, existing experimental procedures only reveal *specific instances of punishment*, and not the full *underlying preference or willingness to engage in punishment*. By contrast, the

advantage of a strategy method design is that it elicits punishment decisions in response to *clean variation across the full range of possible contributions*.

Falk, Fehr and Fischbacher (2005) apply the strategy method to punishment decisions in a one-shot three-person prisoners' dilemma in which the first stage is a binary decision to either cooperate or defect. In this game, there are only four possible combinations of the actions of the other two players. Two of these are symmetric ((*Cooperate, Defect*) and (*Defect, Cooperate*)), but punishment decisions are nonetheless elicited for both cases. The analysis is largely framed in terms of the actions of the punisher and target, and the paper does not report a complete analysis of how punishment varies with the action of the third player. In the main ("high sanction") treatment, it is found that cooperators direct their punishment primarily toward defectors, but that defectors punish both cooperators and other defectors with roughly equal severity. However, one unusual feature of the design is that the "effectiveness of punishment" varies within treatment, depending upon whether punishment is directed toward a cooperator or defector. In particular, each deduction point assigned to a cooperator reduces the earnings of the target by more than one assigned to a defector. Given that the demand for punishment is known to be sensitive to this parameter,<sup>4</sup> it seems likely that this design could distort defectors' relative willingness to punish cooperators versus other defectors.

To redress the limitations of previous research, this paper introduces a simplified three-person voluntary contribution mechanism in which a subject's complete conditional contribution strategy (in a game without punishment) or punishment strategy (in a game with punishment) can be elicited using only ten sets of contingent decisions. This design is used as the framework within which to apply the strategy method in one-shot games both without and with punishment. A within-subjects experiment is conducted, in which subjects play both one-shot games before receiving any feedback on the outcomes of their decisions. As a result of this procedure, each subject can be treated as an independent observation with respect to his or her decisions in both games. As a further contribution, the paper also introduces improved procedures for belief elicitation in public good experiments, including the elicitation of beliefs regarding the severity of punishment that subjects expect to incur from others.

The results of the experiment indicate clearly that behaviours motivated by social preferences are responsive not only to the average level of others' contributions – as has been widely

---

<sup>4</sup> On this point, see Anderson and Putterman (2006), Carpenter (2007) and Nikiforakis and Normann (2008).

presumed in the past<sup>5</sup> – but also to the individual contributions that make up the average. This is the case both for conditional cooperation decisions in the game without punishment, and for conditional punishment decisions in the game with punishment. In this important respect, the findings from the two games are complementary to one another.

In the game without punishment, it is found that there are two distinct sources of selfish bias in conditional cooperation. Firstly, in cases in which the other players contribute equally, the finding of FGF that even subjects classified as conditional cooperators fall short of matching others' contributions is replicated. Secondly, holding the average contributions of the other players constant, conditional contributions decline even further in cases in which the other players contribute more unequally. Further, it turns out that the second form of selfish bias is most pronounced among those for whom the first form is mildest. That is, those who condition more strongly upon others' contributions when the others contribute equally are also more discouraged when the others contribute unequally.

In the game with punishment, half of all subjects are willing to punish in at least one contingency, even in a one-shot interaction. This finding goes some way toward allaying the concern that the strategy method might fail to detect punishment if it weakens the negative emotional response to defection that is thought to trigger acts of punishment (Brandts and Charness, in press). The level of contribution is significantly higher in the game in which punishment is available compared to the one in which it is not.

Most importantly, the elicited punishment function reveals that the comparative statics of punishment respond in a plausible manner to the contributions of both the target and the third player. There is a very substantial positive response in the severity of punishment to negative deviations in the contribution of the target player below that of the punisher, and a smaller negative response to positive deviations above the contribution of the punisher. Further, holding the contribution of the target player constant, there is also a positive response of punishment to the contribution of the third player. All three of these effects are highly statistically significant, and once they are controlled for the contribution of the punisher has no significant effect upon the level of punishment. However, because the response to negative deviations below the punisher is several times stronger than the other two effects, the

---

<sup>5</sup> In the context of conditional cooperation, as already noted, FGF only elicit contributions as a function of the average of others' contributions. In the context of punishment, Fehr and Gächter (2000) analyse the severity of punishment as a function of the deviation of the target player from the average contribution of others.



punishment behaviour of the highest contributors displays the greatest overall responsiveness while the punishment function of the lowest contributors is relatively flat in comparison.

Related to the previous observations, the strategy method also detects “antisocial punishment” (Herrmann, Thöni and Gächter, 2008), where punishment is directed toward a target who contributes at least as much as the punisher. Moreover, the one-shot nature of the interaction precludes many potential explanations for such behaviour – such as retaliation for punishment incurred in the past, or strategic signalling that future punishment will not be tolerated – that arise in repeated games. Contrary to the conjecture that antisocial punishments may be an expression of disdain directed specifically toward “do-gooders”,<sup>6</sup> the results indicate that, if anything, antisocial punishments are slightly *decreasing* in the contribution of the target.

Examining subjects’ beliefs regarding the severity of punishment they expect to incur from others, it is found that non-contributors who punish antisocially expect to incur nearly three times as much punishment compared to non-contributors who do not punish. This finding suggests an explanation for antisocial punishment as a form of pre-emptive retaliation against the punishment that subjects expect to themselves incur from others.

## **DESIGN**

### **Overview**

In the experiments reported in this paper, the basic decision environment is a linear public good game in which there are  $n = 3$  players in each group, and the marginal per capital return on contributions to the public good is  $a = 0.5$ . Each player has an endowment of  $y = 6$  “points” and can choose one of four possible levels of contribution to the public good:  $c \in \{0, 2, 4, 6\}$ . One can interpret these loosely as strong free-riding, weak free-riding, weak cooperation, and strong cooperation, respectively. Each point not contributed generates a private return of 1 earnings point to the individual player only. Each point contributed to the public good by any player generates a return of  $a$  points to each of the  $n$  group members. Since  $n \cdot a > 1 > a$ , full contribution is socially efficient, but for given contributions by the others, an individual’s earnings are always maximised when he or she contributes nothing.

---

<sup>6</sup> “Low contributors might also view high contributors as do-gooders who have shown them up. Punishment may therefore be an act of ‘do-gooder derogation’” (Herrmann, Thöni and Gächter, 2008, p. 1366).

For each subject, there are ten possible ordered combinations of the contributions of the other two group members, and these are the ten cases to which each subject must respond under the strategy method. The contributions of the other players are always presented in the order of lowest followed by highest, and each subject is only presented with the ten cases that are unique under this ordering convention, namely (0, 0), (0, 2), (0, 4), (0, 6), (2, 2), (2, 4), (2, 6), (4, 4), (4, 6), and (6, 6). Thus, for example, it is not necessary to respond separately to the cases (0, 2) and (2, 0).

In the game without punishment, each subject responds by specifying his or her own conditional contribution in each of these ten cases, given the contributions of the others. In the game with punishment, each subject must specify the amount of punishment, if any, that he or she wishes to assign to each of the other two players in each of the cases.<sup>7</sup>

Each subject in the experiment completes both the game without punishment and the game with punishment, where both are played as one-shot games in the strategy method. The order in which the games are played is counterbalanced across sessions, so as to allow the effect of either introducing or removing the punishment opportunity to be examined separately. Subjects are told that they will play two games, and that they will be paid for their decisions in both of them, but they are not told anything about the second game until after they have completed the first one.

Importantly, subjects do not receive feedback on the outcomes of their decisions in either game until after they have completed the second one. In particular, since they do not learn anything about the decisions of others in the first game until after they have made their decisions in the second one, each subject can be treated as an independent observation with respect to his or her decisions in both games. Subjects are told at the start of the second game that they will be matched into a new group of three players, and that they will never be matched with another subject twice in both games.

The use of a one-shot design (together with a perfect strangers matching across games) makes it possible to isolate preferences toward cooperation and punishment in the absence of any strategic or reputational considerations. In such a setting, there is no possible future gain that

---

<sup>7</sup> In the four cases in which the contributions of the other group members are equal, the amount of punishment that is assigned to each of them is not required to be the same.

can ever accrue, either directly or indirectly, from contributing to the public good in the game without punishment, or from inflicting costly punishment in the game with punishment.

The instructions for both games use a standard neutral framing, in which the public good is referred to as “contribution to a project”, and punishment is described as the assignment of “deduction points”. The instructions make no reference to public goods or punishment.<sup>8</sup>

### **The game without punishment**

The game without punishment is an extension of the one-shot strategy method game as introduced by FGF, allowing explicitly for cases in which the other group members may contribute more or less equally. Each subject makes two sets of decisions: a simple “unconditional” contribution decision, and a “contribution table” in which he or she specifies a set of contributions conditional on the decisions of others. Afterwards, one player from each group is randomly chosen to have his or her contribution determined by the contribution table, whereas for the other two players the unconditional contribution is binding. With the contributions thus determined, the earnings of player  $i$  are then given by:

$$\pi_i^N = (y - c_i^N) + a \cdot \sum_{j=1}^n c_j^N$$

where the superscript  $N$  denotes the game without punishment.

In FGF, the cells in the contribution table correspond to each possible *average contribution* of the other group members (rounded to the nearest integer). In the present study, conditional contributions are elicited contingent upon every possible *combination of contributions* of the other group members. In this way, it is possible to compare conditional contributions across pairs of cases in which the average contribution of the other two players is held constant, while the degree to which they contribute more or less equally is varied, for example from (4, 4) to (2, 6). Figure A1 in Appendix A depicts the decision screen for the contribution table of the strategy method game without punishment.

---

<sup>8</sup> Appendix B contains the complete text of the instructions for the “N-P” treatment order in which the game without punishment is played first.

## The game with punishment

In the game with punishment, each subject first makes a simple (unconditional) contribution decision – which need not be the same as in the game without punishment – and there is no contribution table. Thus contributions in the game with punishment are never contingent upon the choices of others, and once a subject has submitted his or her contribution decision it is binding and can no longer be revised.

Each subject can then assign  $p \in \{0, 1, 2, 3\}$  punishment points to each of the other two group members. Each punishment point costs the punisher one earnings point, and reduces the earnings of the target player by  $e = 3$  points. This is subject to the proviso that the punishment that a player incurs from others cannot drive his or her earnings below zero. Nonetheless, it is possible for a subject's earnings to become negative as a result of the cost of the punishment that he or she assigns to others. In this event, the losses are deducted from a “starting balance” of three points that is given to each subject at the very beginning of the session.<sup>9</sup>

At the time they make their punishment decisions, subjects do not know the actual contributions of the other group members. Instead, they are asked to make contingent punishment decisions for each of ten cases, corresponding to the ten possible ordered combinations of the other players' contributions. Each case is presented on a separate screen, but there are “Back” and “Next” buttons available to enable subjects to navigate between the ten cases, and to review and revise their punishment decisions prior to confirming them.

In each of the ten cases, the decision screen displays the subject's own actual contribution, the contributions of the other group members in the case under consideration, and the (hypothetical) resulting earnings for each group member from the contribution stage (before punishment). Figure A2 in Appendix A depicts one of the decision screens for the punishment stage of the strategy method game with punishment.

At the end of the experiment, the computer looks up the actual contributions of the other two players with whom a subject has been matched, to determine which of the ten cases is applicable. Punishment points are only actually allocated for this case. To assign punishment to the other group members, the computer looks up the number of points that the subject

---

<sup>9</sup> There were 11 out of 123 subjects for whom the value of the punishment received exceeded their earnings from the contribution stage. There were 12 out of 123 subjects who went into a loss as a result of assigning punishment to others.

elected to assign in the applicable case. To determine the number of punishment points assigned to a subject, the same is done for the other group members. Given these punishment decisions, the earnings of player  $i$  are then given by:

$$\pi_i^P = \max \left\{ \left[ (y - c_i^P) + a \cdot \sum_{j=1}^n c_j^P - e \cdot \sum_{j \neq i} p_{ji} \right], 0 \right\} - \sum_{j \neq i} p_{ij}$$

where the superscript  $P$  denotes the game with punishment, and  $p_{ij}$  is the number of punishment points assigned by  $i$  to  $j$ .

### **Elicitation of beliefs**

At the end of each game, each subject's beliefs are elicited regarding the contributions of the other subjects in the session. In previous public good experiments, beliefs have only been elicited over the *average or total contribution of the matched group members*.<sup>10</sup> In the present study, beliefs are elicited over the *entire distribution of contributions of all other subjects in the session*.<sup>11</sup> This is done by asking in how many cases out of 100 a subject expects that the other subjects in the session contribute 0, 2, 4, and 6 points. These beliefs are elicited using an incentive-compatible quadratic scoring rule.<sup>12</sup>

In each of the games, each subject can earn up to two additional earnings points depending upon the accuracy of his or her reported beliefs. The instructions explain carefully that a subject's earnings are higher the closer his or her reported beliefs are to the actual distribution of contributions of the other subjects in the session, and that it is not possible for a subject's earnings to be reduced as a result of his or her responses in this task.

In the game with punishment, beliefs are also elicited regarding the total number of deduction points that a subject expects to receive from the other group members. Each subject can earn up to one additional earnings point depending upon the accuracy of this estimate.

---

<sup>10</sup> See for example Fischbacher and Gächter (2010), Gächter and Herrmann (2009), Gächter and Renner (2010), and Neugebauer et al (2009).

<sup>11</sup> Blanco et al (2010) note that the correlation between earnings from the game and earnings from beliefs is reduced when beliefs are elicited over the whole set of other players instead of just the matched players. This reduces the incentive to use stated beliefs as a hedge against adverse outcomes of decisions in the game, and may also make the possibility of hedging less prominent.

<sup>12</sup> See Rey-Biel (2009) and the references contained in Artinger et al (2010) for related applications of the quadratic scoring rule in "multiple choice" settings.

In the context of a one-shot design in which there is no feedback before the end of the session, there is no opportunity for subjects to learn about the behaviour of others while the experiment is in progress. This means that the reported beliefs must derive from the subjects' own introspection based upon life experience outside of the laboratory.

### **Procedures and details of sessions**

Given the one-shot nature of the design, it is essential to take care that subjects have full understanding of the decision problem, so as to ensure that the results are not driven by confusion. To this end, subjects were given ample time in which to read the instructions carefully at their own pace, and to ask any questions privately to the experimenter. Each game did not begin until all subjects in the session correctly answered an extensive set of control questions. There were ten questions relating to payoffs in the absence of punishment, and a further five questions relating to punishment. There was no time limit for subjects to complete these questions. Before the start of each game, the experimenter read aloud a summary of the instructions to ensure that all payoff-relevant information was common knowledge. Finally, there was no time limit for subjects to enter any of their decisions in either game.

The experiments took place in the experimental economics laboratory of an Australian research university in March 2010. A total of 60 subjects took part in three "N-P" sessions in which the game without punishment was followed by the game with punishment. A further 63 subjects took part in three "P-N" sessions in which the treatment order was reversed. Each session involved between 18 and 24 subjects.

Subjects were drawn from a pool of nearly 2,000 undergraduate and postgraduate students from across all fields of study, who had expressed an interest in participating in experiments. For this study, it was decided to recruit only subjects who had never previously participated in any other experiment. A total of 13 out of 123 subjects indicated that they knew one other subject in their session; no-one reported knowing more than one. A total of 61 of the subjects were female, and 62 were enrolled in degrees in Economics and Business. Of the latter, 14 indicated that their major was in Economics.

The average duration of each session was 90 minutes, and the average payment was AUD 28.3 (approximately USD 26.0 or EUR 19.4 at the time the experiments were conducted). The experiment was programmed using z-Tree (Fischbacher, 2007) and the recruitment of subjects was managed using ORSEE (Greiner, 2004).

## RESULTS OF THE GAME WITHOUT PUNISHMENT

In the game without punishment, the primary object of interest is the contribution table, which reveals how a subject's conditional contribution varies as a function of all possible combinations of the contributions of the other two players. As such, the greater part of this section is taken up with an analysis of the properties of conditional contributions. A brief discussion of the other aspects of the game without punishment – namely unconditional contributions and beliefs over the unconditional contributions of others – follows toward the end of this section.

As a preliminary matter before proceeding with the substance of the analysis, it is necessary to check for any possible effect of treatment order. There are ten entries in the conditional contribution table, corresponding to the ten combinations of contributions of the other group members. In each of these ten cells, a Wilcoxon rank-sum test cannot reject the null hypothesis that the distribution of conditional contributions is the same across the two treatment orders ( $p \geq 0.265$ ). Accordingly, the conditional contribution data from both treatment orders can be pooled for the purpose of the analysis that follows.<sup>13</sup>

Next, given that it is well-known that individual behaviour is rather heterogeneous in games of this kind, it is instructive to classify subjects into one of several types on the basis of their conditional contribution behaviour. In particular, a substantial number of studies, using a variety of experimental designs, find that the majority of subjects can be classified either as free-riders or conditional cooperators – where the exact proportions of these two types vary somewhat across populations – with a residual group classified as “others”.<sup>14</sup>

In the analysis that follows, it will be of particular interest to examine how subjects who are classified as “conditional cooperators” respond to increasing inequality in the contributions of others. To avoid biasing this inference, it is therefore appropriate to set aside information on responses to unequal combinations of contributions when deciding who should be classified as a conditional cooperator in the first place. For this reason, the definition of a conditional cooperator is based solely upon a subject's responses along the “diagonal” of the conditional

---

<sup>13</sup> Similarly, there is no evidence of any order effect in the total number of cells in which a subject makes a nonzero contribution ( $p = 0.397$ ), or in the total amount that a subject contributes in aggregate across all ten cells of the conditional contribution table ( $p = 0.326$ , both in Wilcoxon rank-sum tests).

<sup>14</sup> In addition to FGF, see for example Bardsley and Moffatt (2007), Burlando and Guala (2005), Fischbacher and Gächter (2010), Herrmann and Thöni (2009), Kocher et al (2008), Kurzban and Houser (2005), and Muller et al (2008). Much of this literature is surveyed in Chaudhuri (2011).

contribution table, i.e. the cases (0, 0), (2, 2), (4, 4) and (6, 6). This is also consistent with the approach adopted by FGF, who do not elicit any information about “off-diagonal” behaviour.

In particular, a subject is classified as a conditional cooperator if his or her conditional contributions are weakly monotonically increasing along this “diagonal”; that is, if  $c(0, 0) \leq c(2, 2) \leq c(4, 4) \leq c(6, 6)$ , with  $c(0, 0) < c(6, 6)$ , where  $c(l, h)$  denotes a subject’s conditional contribution in response to the case in which the ordered contributions of the other players are  $l$  and  $h$ . This is the case for 41 subjects (33 percent). A subject is classified as a free-rider if he or she enters zero in all ten cells of the contribution table. This is the case for 61 subjects (50 percent). The remaining 21 subjects (17 percent) do not meet either of these criteria, and are classified as “others”.<sup>15</sup> Thus, consistent with previous studies, free-riders and conditional cooperators are the two largest groups, and together these two types account for the vast majority of the sample.<sup>16</sup>

The two panels in Figure 1 depict the mean conditional contributions of the subjects who are classified as conditional cooperators and others, respectively. Each point represents one of the ten cases in the contribution table. These are plotted against the implied mean contribution of the other players on the horizontal axis. Points along the diagonal in this figure correspond to a position of perfect conditional cooperation, in which the mean contributions of others are fully matched. It can be seen from the right-hand panel that the contributions of subjects classified as “others” conform on average to the “hump-shaped” pattern identified by FGF.

In the analysis that follows, it will be instructive to explore the robustness of the results to the definition of a conditional cooperator. To that end, the conditional cooperators will be compared to two alternative classifications of contribution behaviour. For a narrower classification, define a “strong conditional cooperator” as someone for whom at least two of the conditions  $c(0, 0) < c(2, 2)$ ,  $c(2, 2) < c(4, 4)$ , and  $c(4, 4) < c(6, 6)$  hold as strict inequalities, with the third inequality holding at least weakly. There are 33 subjects (27 percent) for whom this is the case. Alternatively, for a broader classification, all 62 subjects (50 percent) who make at least one nonzero entry in the contribution table can be

---

<sup>15</sup> There is no significant relationship between the proportions of subjects classified as free-riders, conditional contributors and “others”, and the order in which the games are played ( $p = 0.384$  in a Pearson chi-square test with two degrees of freedom).

<sup>16</sup> By way of comparison, the proportions identified by FGF in their Swiss subject pool are free-riders 30% and conditional cooperators 50%. FGF define a conditional cooperator as someone whose contribution schedule is weakly monotonically increasing or, failing that, whose Spearman rank correlation coefficient between own and others’ average contributions is positive and significant at the 1% level.



pooled to form the group “conditional cooperators and others”. The mean conditional contributions of each of these two comparison groups are shown in Figure 2. As might be expected, compared to conditional cooperators, the contributions of the more narrowly-defined group lie closer to the diagonal while the contributions of the more broadly-defined group lie further below it. However it is also clear that the qualitative shapes of the functions are quite similar for all three classifications.

The left-hand panel in Figure 1 indicates clearly that there is a selfish bias in the contribution behaviour of conditional cooperators, in that they do not on average fully match the mean contributions of others. Moreover, the figure also clearly highlights the fact that there are two distinct sources of this bias. Firstly, even in the cases (2, 2), (4, 4) and (6, 6) in which the other players contribute equally, conditional cooperators on average fail to match that level of contribution. In effect, this first result replicates the selfish bias identified by FGF.

Secondly, the data also includes three pairs of cases in which the mean of the other players’ contributions are the same, but where in one case these contributions are more unequally distributed than in the other. On all three occasions, it can be seen that the mean contribution of conditional cooperators is further depressed in the case in which the others contribute more unequally. This second effect could not be detected in FGF because they only elicit conditional contributions as a function of the mean of the other players’ contributions.

Table 1 reports nonparametric tests to assess the significance of each of these two sources of selfish bias, not only for conditional cooperators but also the two comparison groups defined above. The top part of the table relates to selfish bias in cases where the other players contribute equally. For example, among conditional cooperators, the mean of  $c(2, 2)$  is 1.610. A Wilcoxon signed-rank test rejects the null hypothesis that  $c(2, 2) = 2$  with a  $p$ -value of 0.033. Likewise, for conditional cooperators,  $c(4, 4)$  and  $c(6, 6)$  differ significantly from 4 and 6 respectively, with  $p < 0.001$  in signed-rank tests in both cases.

As might be expected, this first form of selfish bias is milder when the stronger definition of a conditional cooperator is adopted, and more pronounced when the broader classification is considered instead. In particular, among strong conditional cooperators it is not possible to reject the null hypothesis that  $c(2, 2) = 2$  ( $p = 0.257$ ), although it remains the case that  $c(4, 4)$  and  $c(6, 6)$  differ significantly from 4 and 6 respectively ( $p \leq 0.002$ , all in signed-rank tests).

The bottom part of Table 1 reports  $p$ -values for Wilcoxon signed-rank tests of the increase in selfish bias that accompanies an increase in the inequality in the other players' contributions, holding constant the mean. For all three pairs of cases for which this comparison can be made, and for all three groupings of subjects, the table confirms that contributions are significantly lower (with  $p \leq 0.024$ ) when the other players' contribute more unequally. This finding confirms that previous designs overlook important information when they elicit conditional contributions solely as a function of the mean of others' contributions – clearly, the degree of inequality matters a great deal as well.

As it turns out, this second form of selfish bias is most severe when the narrower type classification is considered, and weakest when the broader version is adopted instead. In other words, subjects who condition their own contributions more strongly upon the contributions of others, in cases where others contribute equally, are also more discouraged when the others contribute unequally. The right-hand panel in Figure 1 shows that subjects classified as “others” do not respond as negatively to increasing inequality in the contributions of others.

One limitation of the nonparametric tests reported in Table 1 is that they do not exploit the full set of ten conditional contribution decisions made by each subject, instead relying upon pairwise comparisons. To provide a more complete account of the two sources of selfish bias, Table 2 reports random effects interval regressions in which each conditional contribution decision is regressed on the implied mean and difference in the other players' contributions in the decision case that generated it, a dummy for the treatment order, and a constant. These regressions are reported for the conditional cooperators, as well as the two comparison groups defined above. Each subject contributes ten observations to the regression, corresponding to his or her decisions in each of the ten cases in the conditional contribution table.

A random effects specification is employed to capture unobserved individual-specific heterogeneity in the propensity to contribute. An interval regression model is used on account of the discrete nature of the contribution space, which is limited to one of only four possible contribution levels, causing GLS or Tobit errors to be heteroskedastic. Out of these four permissible contributions, a subject is simply assumed to choose the one that lies closest to his or her “true” desired contribution. Thus an observed contribution of 0 is taken to imply that the desired contribution is less than 1, an observed contribution of 2 is taken to imply that the

desired contribution must lie between 1 and 3, and so on. Table 2 also reports marginal effects for these regressions, conditional upon contributions lying in the interval between 0 and 6.<sup>17</sup>

The marginal effects reported in Table 2 imply that conditional cooperators respond to a one-point increase in the mean of the other players' contributions – holding the difference between them constant – by increasing their desired contribution by 0.833 points. This response differs significantly from one ( $Z = -4.28$ ,  $p < 0.001$ ). On the other hand, conditional cooperators respond to a one-point increase in the difference between the other players' contributions – holding the mean constant – by *decreasing* their desired contribution by 0.136 points. This response differs significantly from zero ( $p < 0.001$ ).

Turning next to the two comparison groups, for strong conditional cooperators the response to an increase in the mean is closer to one than it is for conditional cooperators, but still differs significantly from one ( $Z = -2.05$ ,  $p = 0.040$ ), while the response to an increase in the difference is more negative than it is for conditional cooperators. This reiterates the earlier observation that those subjects who respond most strongly when the others contribute equally are also more discouraged when the others contribute unequally. For the broader comparison group, consisting of all subjects who make at least one nonzero conditional contribution, it is not possible to reject the null hypothesis that contributions do not respond to the difference between the other players' contributions ( $p = 0.309$ ).

The regression model in Table 2, in which the other players' contributions are entered in the form of the mean and difference, is also equivalent to an alternative specification in which the minimum and maximum contributions are entered directly into the regression – given that there are only two other players, these two pairs of variables are perfectly collinear. Marginal effects for this alternative specification are also reported at the bottom of Table 2.

The results show that conditional cooperators respond to a one-point increase in the contribution of the lower contributor – holding that of the higher contributor constant – by increasing their desired contribution by 0.552 points. On the other hand, conditional cooperators respond to a one-point increase in the contribution of the higher contributor – holding that of the lower contributor constant – by increasing their desired contribution by

---

<sup>17</sup> For these models, the first set of marginal effects, corresponding to the mean and difference of the other players' contributions, are applicable. The second set of marginal effects, for the minimum and maximum of the other players' contributions are discussed further below.

only 0.281 points. The null hypothesis that the responses to the minimum and maximum contributions are equal is soundly rejected with  $p < 0.001$  in a Wald test.

By comparing these estimates to the original version of this regression, it is easy to see why conditional contributors are more responsive to a one-point increase in the minimum of the other players' contributions than they are to a one-point increase in the maximum. Although either increase has the same implied effect upon the mean, an increase in the lower contribution has the effect of *decreasing* the inequality in the other players' contributions, whereas an increase in the higher contribution has the opposite effect. Thus in the first case, the response to decreased inequality reinforces the response to the increased mean, whereas in the second case, the response to increased inequality works in the opposite direction.<sup>18</sup>

Finally, Tables 3 and 4 summarise key results regarding, respectively, unconditional contributions and beliefs regarding the unconditional contributions of others in the game without punishment. Table 3 reports the cross-tabulation of a subject's unconditional contribution with his or her type as defined from the contribution table.<sup>19</sup> The modal unconditional contribution is zero, and three-quarters of those who unconditionally contribute zero are also classified as free-riders on the basis of their contribution table. Among those classified as free-riders, the mean unconditional contribution is close to zero, whereas among conditional cooperators it is close to the midpoint of the contribution space.

Table 4 reports the means of the number of cases out of 100 in which a subject expects that the others in the session will make unconditional contributions of 0, 2, 4, and 6, as well as the implied expectation of other subjects' contributions, disaggregated according to a subject's own unconditional contribution and conditional contribution type.<sup>20</sup> In aggregate, the modal belief is that others will contribute zero. Subjects who contribute zero do so even though they expect that others will make a nonzero contribution roughly one-third of the time. On the

---

<sup>18</sup> To see how the marginal effects from the two specifications relate to one another, note that a one-point increase in the minimum increases the mean by half a point, while decreasing the difference by one point. Thus according to the original model, the predicted response is  $0.833/2 + 0.136 = 0.552$ . On the other hand, a one-point increase in the maximum increases the mean by half a point, while *increasing* the difference by one point. Thus according to the original model, the predicted response is  $0.833/2 - 0.136 = 0.281$ . Finally, turning briefly to the comparison groups, for the broader group consisting of all who make at least one non-zero contribution, there was no significant response to the difference in others' contributions. Accordingly, for this group there is also no significant difference between the response to the minimum and maximum.

<sup>19</sup> There is no significant effect of treatment order upon unconditional contributions in the game without punishment ( $p = 0.545$  in a Wilcoxon rank-sum test).

<sup>20</sup> There is no significant effect of treatment order on subjects' implied expectation regarding the unconditional contributions of others in the game without punishment ( $p = 0.272$  in a Wilcoxon rank-sum test).

other hand, subjects who make unconditional contributions of 4 and 6 do so in spite of the fact that they do not expect these contributions to be matched by others. The beliefs of conditional cooperators are significantly more optimistic than those of free-riders ( $p < 0.001$  in a Wilcoxon rank-sum test on the implied mean contribution belief).

## RESULTS OF THE GAME WITH PUNISHMENT

The analysis of the game with punishment is complicated somewhat by evidence of an order effect. The level of contributions in the game with punishment is significantly higher when this game is played first compared to when it is played after the game without punishment ( $p = 0.015$  in a Wilcoxon rank-sum test). Figure 3 summarises the mean levels of (unconditional) contribution in each treatment and order. While contributions are significantly higher when punishment is available compared to when it is not in both treatment orders,<sup>21</sup> it appears that the impact of punishment is attenuated when subjects have previous experience of the game without punishment. Since subjects receive no feedback until the end of the session, this effect cannot be attributed to learning about the behaviour of others.

The same order effect is also evident in the propensity to punish. In the P-N treatment order, 37 out of 63 subjects (59 percent) assign nonzero punishment on at least one occasion. In the N-P treatment order, the corresponding number is 24 out of 60 (40 percent). A Pearson chi-square test with one degree of freedom confirms that there is a significant association between willingness to punish and treatment order ( $p = 0.038$ ). Thus in aggregate, half of the subjects are willing to punish on at least one occasion. This confirms that the strategy method is indeed capable of detecting considerable willingness to punish, notwithstanding the fact that it may diminish the emotional impact of the interaction (Brandts and Charness, in press).

*Conditional upon willingness to punish*, there little evidence of any order effect in the *severity of punishment*. Among those subjects who punish at least once, and in all but one of the twenty conditional punishment decisions, there is no significant order effect in the number of punishment points assigned ( $p \geq 0.106$  in Wilcoxon rank-sum tests).<sup>22</sup> For this reason, the punishment data from the two treatment orders will be pooled for the purposes of graphical presentation. This is done on the grounds that it is the *comparative statics of punishment* that are of primary interest, and there is little evidence to suggest that these might be affected by

---

<sup>21</sup> For the N-P treatment order  $p = 0.057$ , and for the P-N order  $p < 0.001$ , both in Wilcoxon signed-rank tests.

<sup>22</sup> The exception relates to a case in which both of the other players make the full contribution of 6. In this case only, there is a significant difference by treatment order with  $p = 0.028$  in the rank-sum test.

the treatment order. However, this is also subject to the caveat that the *absolute level of punishment* will depend upon the proportion of subjects who are willing to punish, and this is clearly sensitive to the treatment order. Appendix A contains supplementary figures in which the two treatment orders are presented separately. Regression analyses are reported both for the pooled data (including a dummy for treatment order) and separately for each of the treatment orders.

Figure 4 presents a first look at the punishment data in a format first introduced by Fehr and Gächter (2000). It depicts the average over all punishment decisions of the number of punishment points that a subject assigns to another player, as a function of the deviation in the contribution of the target player from the average contribution of the other group members (i.e. the punisher and the third player). This figure shows that punishment decisions elicited using the strategy method in a one-shot game display the same behavioural regularities that Fehr and Gächter identified in repeated games using the direct-response method. When the contribution of the target deviates below the average of the punisher and third player, the severity of punishment increases substantially in the magnitude of the deviation. When the target player contributes above the average, some punishment continues to occur, but it is of much lesser magnitude and does not appear to respond to the size of the deviation.<sup>23</sup> This replication of one of the major results in Fehr and Gächter (2000) provides an important validation of the use of the strategy method in this setting.

However, the presentation in Figure 4 may not provide the most insightful picture of the forces that influence punishment behaviour. In particular, variation along the horizontal axis is a function of the contributions of all three players. The average contribution is a moving target that depends upon both the contribution of the punisher (which is endogenous) and that of the third player (which varies systematically over the different cases in the strategy method design). As a result, an increase in the negative deviation may arise either from a decrease in the contribution of the target, or an increase in either of the contributions that make up the average. Finally, some values of the deviation arise only from very specific combinations of contributions, while others are more ubiquitous. For example, a deviation of  $-6$  is only observed when both the punisher and third player contribute 6 while the target contributes 0.

---

<sup>23</sup> Figure A3 in Appendix A reports the analysis in Figure 4 separately for each treatment order. Because of the higher propensity to punish in the P-N treatment order, the level of the function is somewhat higher in the right-hand panel. However, with the exception of behaviour at the two extremities, the shape of the function is similar for both treatment orders. Note that because the level of contribution is higher in the P-N order, there are relatively more observations represented in the  $-6$  cell, and fewer in the  $+6$  cell in this order.

Accordingly, the number of such observations is limited to the number of subjects who contribute 6. By contrast, there are multiple combinations of contributions that give rise to deviations close to 0.

Figure 5 provides a clearer picture of how the punishment behaviour of subjects varies across the ten cases (and twenty punishment decisions) elicited under the strategy method. It shows the average over all subjects of the number of punishment points assigned as a function of the contribution of the target, further decomposed by the contribution of the third player.<sup>24</sup> This figure displays two clear regularities. Firstly, holding constant the contribution of the third player, the severity of punishment directed toward the target increases as the contribution of the target falls. Secondly, holding constant the contribution of the target, the severity of punishment also increases as the contribution of the third player rises. Whereas these two forces were conflated in the presentation of Figure 4, it is clear from Figure 5 that they each appear to exert an independent influence.<sup>25</sup>

Because Figure 5 pools the punishment decisions of all subjects, it can be interpreted as a measure of the severity of punishment that a player might expect to incur on average, if he or she were to encounter a randomly-drawn subject from the experiment. However, this presentation masks the differences in punishment behaviour associated with different levels of contribution by the punisher. To shed further light upon this, Figure 6 disaggregates the analysis in Figure 5 by displaying the average punishment functions separately for those who contribute 0, 2, 4, and 6 points respectively.<sup>26</sup>

Figure 6 highlights several related observations. Firstly, even those who behave in a selfish manner by contributing 0 do not necessarily refrain from assigning costly punishment. On the other hand, even those who make the maximum contribution of 6 points do not necessarily escape punishment. Thus the strategy method detects so-called “antisocial punishment” (Herrmann, Thöni and Gächter, 2008). In the discussion that follows, antisocial punishment is defined as where a subject punishes a player whose contribution is at least as great as his or her own. In this event, the effect of punishment is to increase the earnings differential

---

<sup>24</sup> Recall that in the four cases where the contributions of the other players are equal, the punishment assigned to each of them need not be the same. In these cases, both sets of observations are pooled for the purpose of the analysis in Figures 5 and 6.

<sup>25</sup> Figure A4 in Appendix A reports the analysis in Figure 5 separately for each treatment order. Each panel in this figure displays both of the patterns noted in the text. The higher level of the function in the right-hand panel simply reflects the greater propensity to punish under the P-N order.

<sup>26</sup> Figures A5a and A5b in Appendix A report the analysis in Figure 6 separately for each treatment order.

between the punisher and a target whose earnings before punishment are already less than or equal to those of the punisher. It follows from the definition that any nonzero punishment by a subject who contributes 0 is necessarily antisocial and, conversely, that the greater a subject's own contribution, the fewer of his or her punishment decisions are potentially antisocial.

A final observation regarding Figure 6 is that the shape of the punishment function clearly varies substantially across the different levels of contribution of the punisher. Among those who contribute 6 points – for whom the vast majority of punishments are not antisocial – punishment behaviour clearly conforms to the two patterns identified in Figure 5. On the other hand, among those who contribute 0 points – for whom all punishments are by definition antisocial – the punishment function appears to be much flatter.

To organise these findings, Table 5 reports random effects interval regressions in which each conditional punishment decision is regressed upon the contribution of the punisher, the contribution of the third player, the absolute negative deviation in the contribution of the target below that of the punisher, the positive deviation of the target above the punisher, a dummy for treatment order, and a constant.<sup>27</sup> The negative and positive deviations of the target are entered separately to allow for the possibility that antisocial punishments might respond differently to the contribution of the target, compared to punishments of lower contributors. Each subject contributes twenty observations to the regression, corresponding to the punishment they assign to each of two other players in each of ten cases.

The results in Table 5 indicate that, after controlling for the other variables, the contribution of the punisher has no significant effect upon the severity of punishment, while the contribution of the third player has a significant positive effect. Negative deviations in the contribution of the target below that of the punisher have a significant positive effect upon punishment – and the magnitude of this effect is much larger than any of the others. Finally, positive deviations in the contribution of the target above that of the punisher have a negative effect upon punishment, although this is much smaller in magnitude than the response to negative deviations (and in the N-P treatment order it is not statistically significant).<sup>28</sup> Thus

---

<sup>27</sup> The table also reports separate regressions by treatment order. As before, interval regression models are used on account of the discrete nature of the punishment space, which is limited to 0, 1, 2, or 3 points. Out of these four permissible punishments, the subject is again assumed to choose the one that lies closest to his or her “true” desired punishment. Table 5 also reports marginal effects for these regressions, conditional upon punishment lying in the interval between 0 and 3.

<sup>28</sup> The null hypothesis that the coefficients on absolute negative and positive deviations are equal (but opposite in sign) is soundly rejected, with  $p < 0.001$  in a Wald test.



the model in Table 5 does not support the notion that antisocial punishments are targeted specifically toward high contributors.<sup>29</sup>

Table 6 reports the marginal effects (conditional upon punishment lying in the interval between 0 and 3) from an enlarged model in which the contribution of the third player and the absolute negative and positive deviations of the target from the punisher are interacted with dummies for each level of contribution by the punisher. This model therefore allows the responses to the contributions of the other players to differ across different levels of contribution by the punisher. In the model that pools observations from both treatment orders, the response to the contribution of the third player is positive and significant at every contribution level of the punisher.<sup>30</sup> Likewise, the absolute negative deviation of the target below the punisher always has a sizable and significant positive effect upon punishment. For punishers who contribute 2 or 4, the responses to positive deviations of the target above the punisher are not significant in the pooled model.<sup>31</sup> Among those who contribute 0, there is a significant *negative* response of punishment to positive deviations in the contribution of the target. Once again, the sign of this effect is contrary to the suggestion that non-contributors target their antisocial punishments specifically toward high contributors.

Finally, Table 7 reports summary measures of beliefs regarding the total amount of punishment that subjects expect to incur from others, disaggregated by the subject's own contribution, whether or not the subject ever punishes, and whether or not the subject ever punishes antisocially. Two broad facts are evident from this analysis. Firstly, subjects who punish also expect to incur more punishment from others than those who do not. Secondly, among subjects who punish, those who sometimes punish antisocially again expect to incur more punishment than those who do not. Both effects are highly significant in Wilcoxon rank-sum tests when the data are pooled over all levels of contribution of the punisher ( $p \leq 0.004$ ).

---

<sup>29</sup> Herrmann, Thöni and Gächter (2008, p. 1366) in fact report likewise that, in the majority of their subject pools, antisocial punishment was lower the higher the contribution of the target. They interpret this finding to suggest that "(s)ome antisocial punishment may be efficiency-enhancing in intent", however they do not acknowledge its apparent contradiction with the conjecture of "do-gooder derogation".

<sup>30</sup> However, in the N-P treatment order it is insignificant when the punisher contributes 6, and in the P-N treatment order it is insignificant when the punisher contributes 0. In the latter case, the lack of significance seems likely to stem in part from the fact that there are fewer subjects who contribute 0 in the P-N order.

<sup>31</sup> However, in the N-P treatment order the effect is positive and significant when the punisher contributes 4 (and also marginally significant when the punisher contributes 2). In the P-N treatment order the effect is negative and significant when the punisher contributes 2.

Subjects who contribute 0 are of particular interest since for this group all punishments are by definition antisocial. Subjects who contribute 0, but who do not punish, expect on average that they will incur 1 punishment point. On the other hand, subjects who contribute 0 and who also punish antisocially expect on average to incur 2.9 punishment points. This difference is highly significant, with  $p = 0.002$  in a Wilcoxon rank-sum test. This result suggests an interpretation of antisocial punishment as a form of *pre-emptive retaliation* against the punishment that low contributors expect to incur from others.

## CONCLUSION

Through the application of the strategy method, this paper makes several contributions toward understanding behaviour in voluntary contribution experiments both without and with punishment. In a game without punishment, it is found that conditional contributions are responsive not only to the average contribution of other players, but also the extent to which they contribute more or less equally. It is thus shown that there are two distinct sources of selfish bias in the contributions of conditional cooperators. Firstly, they do not match the contributions of other players, even when the others contribute equally. Secondly, for a given average contribution of the other players, conditional cooperators reduce their contributions even further as others contribute more unequally.

In a game with punishment, it is demonstrated that the strategy method can indeed detect contingent willingness to punish. The punishment behaviour elicited under the strategy method varies in a plausible manner both as a function of the contribution of the target player relative to the punisher, and the contribution of the third player. Finally, the strategy method also detects antisocial punishment. Rather than being an expression of disdain targeted specifically toward “do-gooders”, it is found that, if anything, antisocial punishments in fact respond slightly negatively to the contribution of the target. Instead, the beliefs of antisocial punishers suggest that their behaviour may be motivated by pre-emptive retaliation against the punishment they expect to incur from others.

How can these results be interpreted in light of the literature on theories of social preferences? Both conditional cooperation and punishment are commonly interpreted as manifestations of reciprocity, the desire to be kind toward others who have been kind to oneself and unkind toward those who have not. Indeed, in the context of one-shot public good games played in the direct-response mode, Gächter and Herrmann (2009) equate contribution in a game

without punishment to strong positive reciprocity, and punishment of non-contributors in a game with punishment to strong negative reciprocity.

Sugden (1984) proposes a specific form of reciprocity which “says, with certain qualifications, that if everyone else contributes a particular level of effort to the production of a public good, you must do the same” (p. 776). That is, there is a moral obligation to contribute at least the level that matches the minimum of others’ contributions. The “qualification” is that when the level of contribution that one most prefers everyone to make is less than the minimum, one is only obliged to contribute at that level. Bardsley and Moffatt (2007) and Croson (2007) report tests of Sugden’s model in public good experiments, with mixed results. Bardsley and Moffatt find the behaviour of reciprocators to be insufficiently reciprocal relative to Sugden’s benchmark. On the other hand, Croson finds the median of others’ contributions to be a better predictor of individual behaviour than the minimum.

In broad terms, Sugden’s model fits both main stylised facts observed in this paper as regards the game without punishment. The proposition that in general one is only morally obliged to match the minimum of others’ contributions can explain why conditional contributions decline as others contribute more unequally, holding their average contribution constant. At the same time, Sugden’s “qualification” provides the “get-out clause” that permits some subjects to contribute less than others do, even when the others contribute equally.

However, both conditional cooperation and the two forms of selfish bias are also compatible with the intuition of inequality aversion. In particular, the model of Fehr and Schmidt (1999) asserts that, in addition to utility from one’s own material payoff, a subject experiences disutility from inequality in material payoffs, and that disadvantageous inequality is more keenly felt than advantageous inequality. Clearly, the inequality in payoffs is minimised when all contribute equally, and this provides an impetus to increase one’s own contributions in line with others. If the other players contribute equally, a subject who contributes less than they do faces a trade-off between increasing material payoffs and disutility from increasing advantageous inequality. If the coefficient on advantageous inequality is sufficiently small, it will be optimal to lower one’s own contribution below the level of the others.

Next, consider what happens when the other players contribute unequally, while their mean contribution remains unchanged. If one contributes at the same level as before, one now faces a position of advantageous inequality with respect to the higher contributor, and disadvantageous inequality with respect to the lower contributor. Since the latter is more

keenly felt than the former, this provides a further impetus to lower one's own contribution, relative to the initial position in which the others contributed equally. As one lowers one's contribution, not only are material payoffs increased, but the reduction in disutility from disadvantageous inequality with respect to the lower contributor outweighs the increase in disutility from advantageous inequality with respect to the higher contributor. Thus, for a given mean contribution of the other players, if one is more averse to disadvantageous than to advantageous inequality, one should contribute less when the other players contribute unequally compared to when they contribute equally.

Interpreting behaviour in the game with punishment is a more speculative matter. It is clear that punishing a player who contributes less than oneself is consistent either with negatively reciprocating that player's unkind deed, or with reducing disadvantageous inequality relative to that player. In a binary-choice prisoners' dilemma, Falk, Fehr and Fischbacher (2005) find that cooperators continue to punish defectors even when doing so does not affect the earnings differential between them, and therefore inequality aversion cannot be the sole motivation.

The interpretation of the finding that, holding constant the contributions of the punisher and target, the severity of punishment increases in the contribution of the third player does not appear to have been remarked upon previously. One possibility is that the punisher recognises the third player's contribution as a kind deed that deserves to be reciprocated. Given that the players' contribution decisions are already binding, the only instrument available for reciprocation is through punishment. Accordingly, the punisher may reciprocate the third player's contribution by bearing more of the cost of disciplining a lower contributor.

Alternatively, notice that although the third player's contribution does not affect the earnings differential between the punisher and target, it raises the earnings of both relative to the third player. This may increase the attractiveness of punishing a lower-contributing target: not only does punishment reduce the punisher's disadvantageous inequality relative to the target, the cost of punishment also reduces advantageous inequality relative to the third player.

Finally, this paper provides evidence that antisocial punishment may be a form of pre-emptive counter-punishment, and that it does not appear to be consistent with "do-gooder derogation". Nonetheless, some alternative explanations remain open. Firstly and most obviously, antisocial punishment could be a product of spiteful preferences, whereby a punisher derives positive utility from advantageous inequality. In the context of a prisoners' dilemma, Falk, Fehr and Fischbacher (2005) find that the punishment of cooperators by defectors largely

disappears when it cannot affect the earnings differential between them, and conclude from this that such punishments are driven by spite.

Other explanations that have been put forward for antisocial punishment include decision error (Fehr and Gächter 2000) and suspicion toward the generosity of others (Herrmann, Thöni and Gächter 2008). More subtly, a recent paper by Thöni (2011) identifies circumstances in which antisocial punishment may be motivated by inequality aversion. These relate to situations in which a punisher wishes to target a lower-contributing player, but there is a high-contributing third player who is unwilling to punish. In this case, the punisher may elect to punish *both* of the other players. Here, the motive for punishing the high contributor is to avoid falling behind that player's earnings on account of the high contributor's unwillingness to share in the costs of punishment.

Whatever the case may be, it is very much to be hoped that the new experimental designs and results contributed in this paper may help to inform the on-going development of theory, as well as to stimulate further studies that might ultimately be able to discriminate between some of the alternative explanations that have been canvassed above.

## REFERENCES

- Anderson, C.M., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior* 54, 1–24.
- Artinger, F., Exadaktylos, F., Koppel, H., Sääksvuori, L., 2010. Applying quadratic scoring rule transparently in multiple choice settings: A note. Max-Planck-Institute of Economics, Friedrich-Schiller-University Jena.
- Bardsley, N., Moffatt, P.G., 2007. The experimentics of public goods: Inferring motivations from contributions. *Theory and Decision* 62, 161–193.
- Blanco, M., Engelmann, D., Koch, A.K., Normann, H.-T., 2010. Belief elicitation in experiments: Is there a hedging problem? *Experimental Economics* 13, 412–438.
- Brandts, J., Charness, G., in press. The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*.
- Burlando, R.M., Guala, F., 2005. Heterogeneous agents in public goods experiments. *Experimental Economics* 8, 35–54.
- Carpenter, J.P., 2007. The demand for punishment. *Journal of Economic Behavior and Organization* 62, 522–542.
- Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics* 14, 47–83.
- Cooper, D.J., Kagel, J.H., in press. Other regarding preferences: A selective survey of experimental results, in: Kagel, J.H., Roth, A.E. (Eds.), *The Handbook of Experimental Economics*, Volume 2. Princeton University Press, Princeton.
- Croson, R.T.A., 2007. Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. *Economic Inquiry* 45, 199–216.
- Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. *Econometrica* 73, 2017–2030.
- Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90, 980–994.
- Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. *Nature* 415, 137–140.
- Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 817–868.
- Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10, 171–178.
- Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of free riding in public goods experiments *American Economic Review* 100, 541–556.
- Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71, 397–404.

- Gächter, S., Herrmann, B., 2009. Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of The Royal Society B* 364, 791–806.
- Gächter, S., Renner, E., 2010. The effects of (incentivized) belief elicitation in public good experiments. *Experimental Economics* 13, 364–377.
- Greiner, B., 2004. The online recruitment system ORSEE 2.0: A guide for the organization of experiments in economics. Department of Economics, University of Cologne.
- Herrmann, B., Thöni, C., 2009. Measuring conditional cooperation: A replication study in Russia. *Experimental Economics* 12, 87–92.
- Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. *Science* 319, 1362–1367.
- Keser, C., van Winden, F., 2000. Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics* 102, 23–39.
- Kocher, M.G., Cherry, T., Kroll, S., Netzer, R.J., Sutter, M., 2008. Conditional cooperation on three continents. *Economics Letters* 101, 175–178.
- Kurzban, R., Houser, D., 2005. Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences* 102, 1803–1807.
- Ledyard, J.O., 1995. Public goods: A survey of experimental research, in: Kagel, J.H., Roth, A.E. (Eds.), *The Handbook of Experimental Economics*. Princeton University Press, Princeton, pp. 111–194.
- Muller, L., Sefton, M., Steinberg, R., Vesterlund, L., 2008. Strategic behavior and learning in repeated voluntary contribution experiments. *Journal of Economic Behavior and Organization* 67, 782–793.
- Neugebauer, T., Perote, J., Schmidt, U., Loos, M., 2009. Selfish-biased conditional cooperation: On the decline of contributions in repeated public goods experiments. *Journal of Economic Psychology* 30, 52–60.
- Nikiforakis, N., Normann, H.-T., 2008. A comparative statics analysis of punishment in public-good experiments. *Experimental Economics* 11, 358–369.
- Rey-Biel, P., 2009. Equilibrium play and best response to (stated) beliefs in normal form games. *Games and Economic Behavior* 65, 572–585.
- Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopol-experiments, in: Sauermann, H. (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung*. J.C.B. Mohr (Siebeck), Tübingen, pp. 136–168.
- Sugden, R., 1984. Reciprocity: The supply of public goods through voluntary contributions. *Economic Journal* 94, 772–787.
- Thöni, C., 2011. Inequality aversion and antisocial punishment. University of St. Gallen.

Figure 1. Mean conditional contributions as a function of combinations of contributions of the other two players, for subjects classified as “conditional cooperators” ( $n = 41$ ) and “others” ( $n = 21$ ). The horizontal axis depicts the mean of the other two players’ contributions. The diagonal corresponds to perfect conditional cooperation.

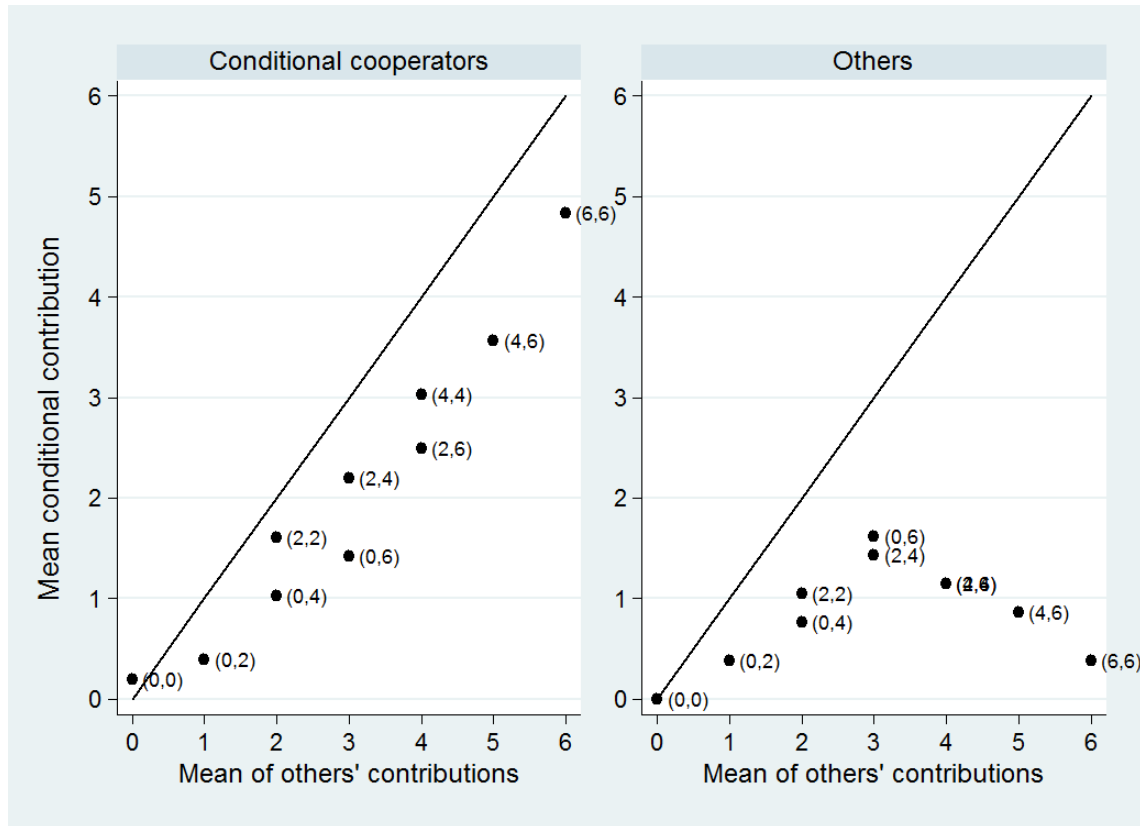




Figure 2. Mean conditional contributions as a function of combinations of contributions of the other two players, for subjects classified as “strong conditional cooperators” ( $n = 33$ ) and “conditional cooperators and others” ( $n = 62$ ). The horizontal axis depicts the mean of the other two players’ contributions. The diagonal corresponds to perfect conditional cooperation.

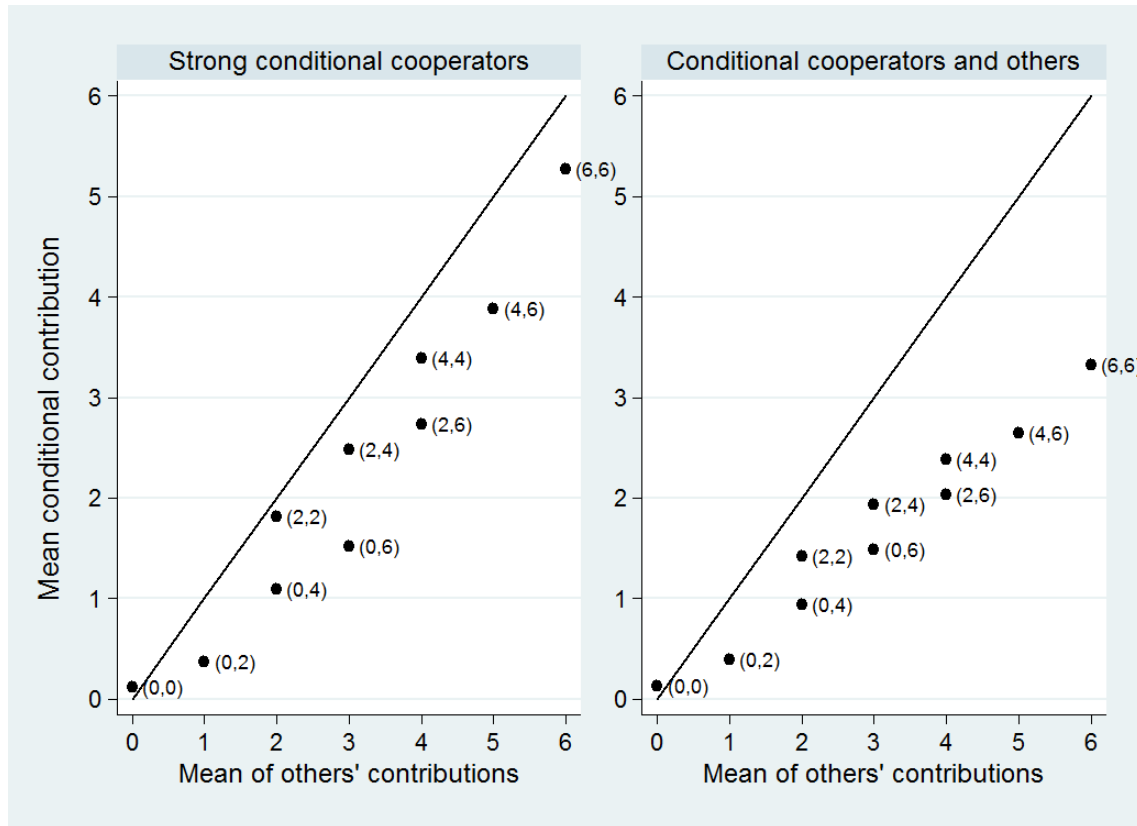


Figure 3. Summary of mean (unconditional) contributions by treatment and treatment order.

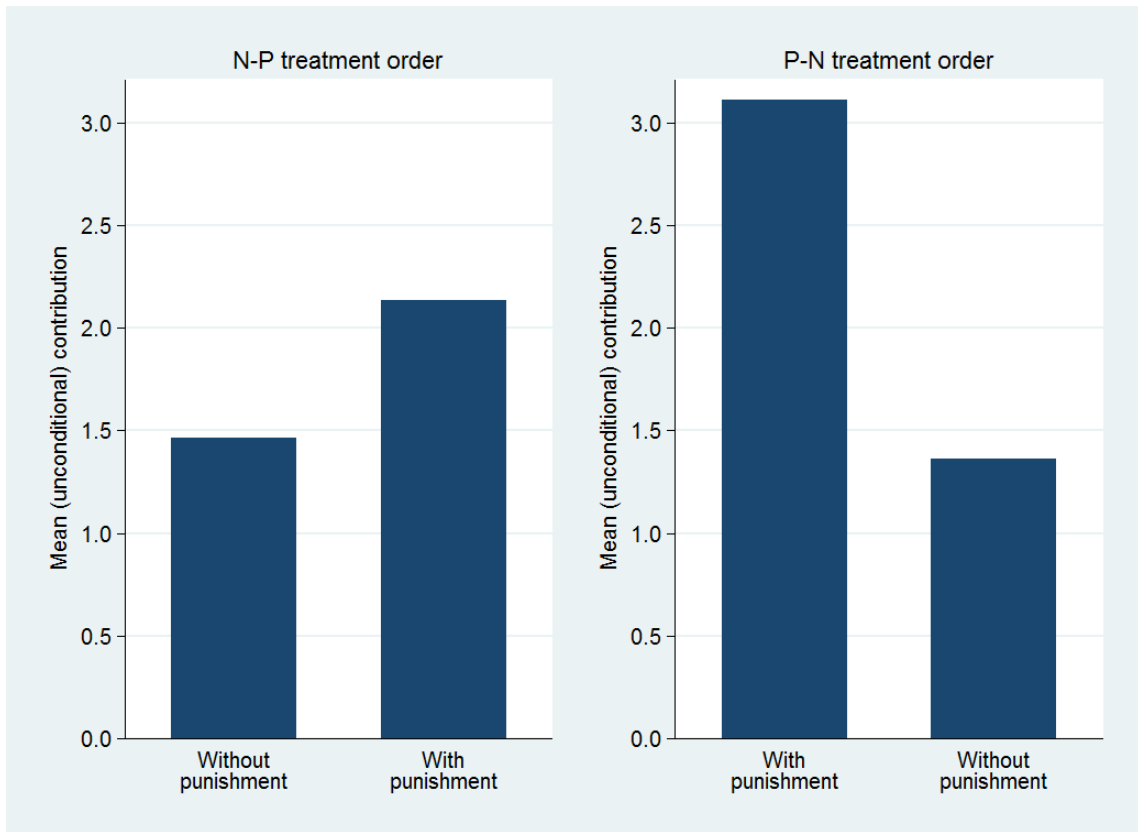


Figure 4. Mean punishment points assigned, as a function of the deviation in the contribution of the target player from the mean contribution of the punisher and third player.

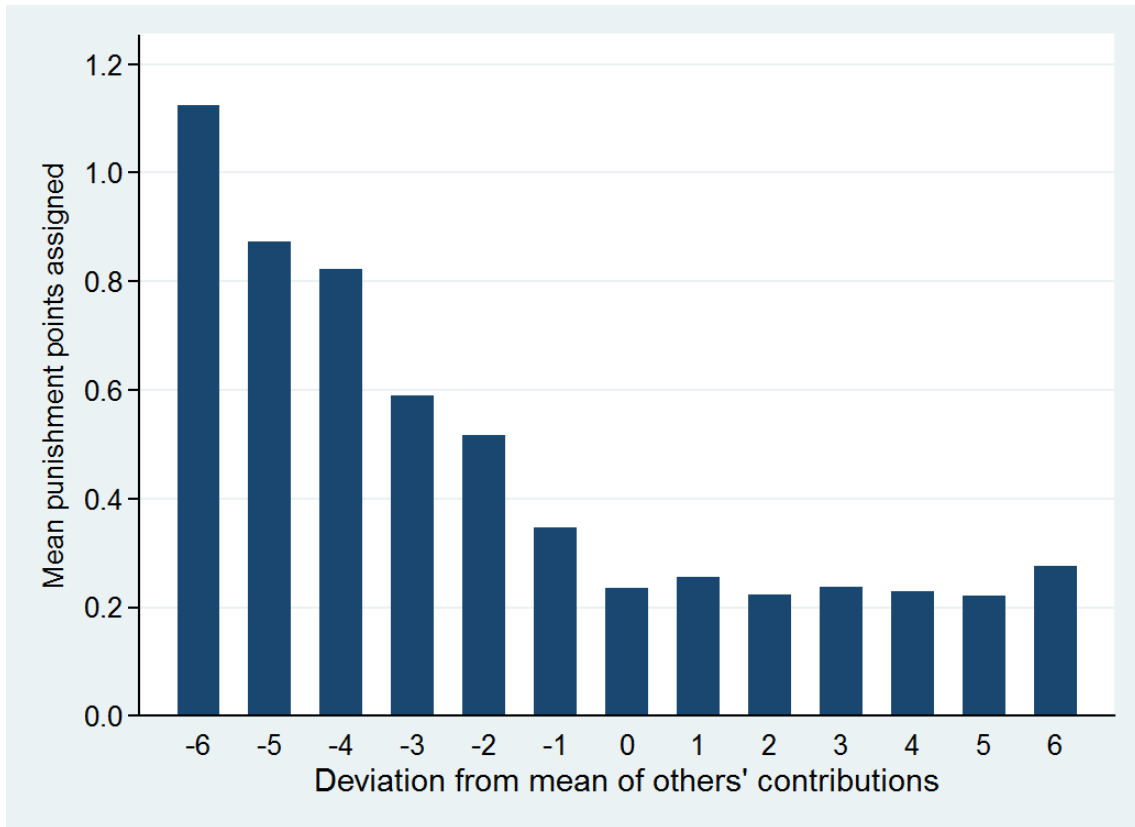


Figure 5. Mean punishment points assigned, as a function of the contributions of the target player and third player.

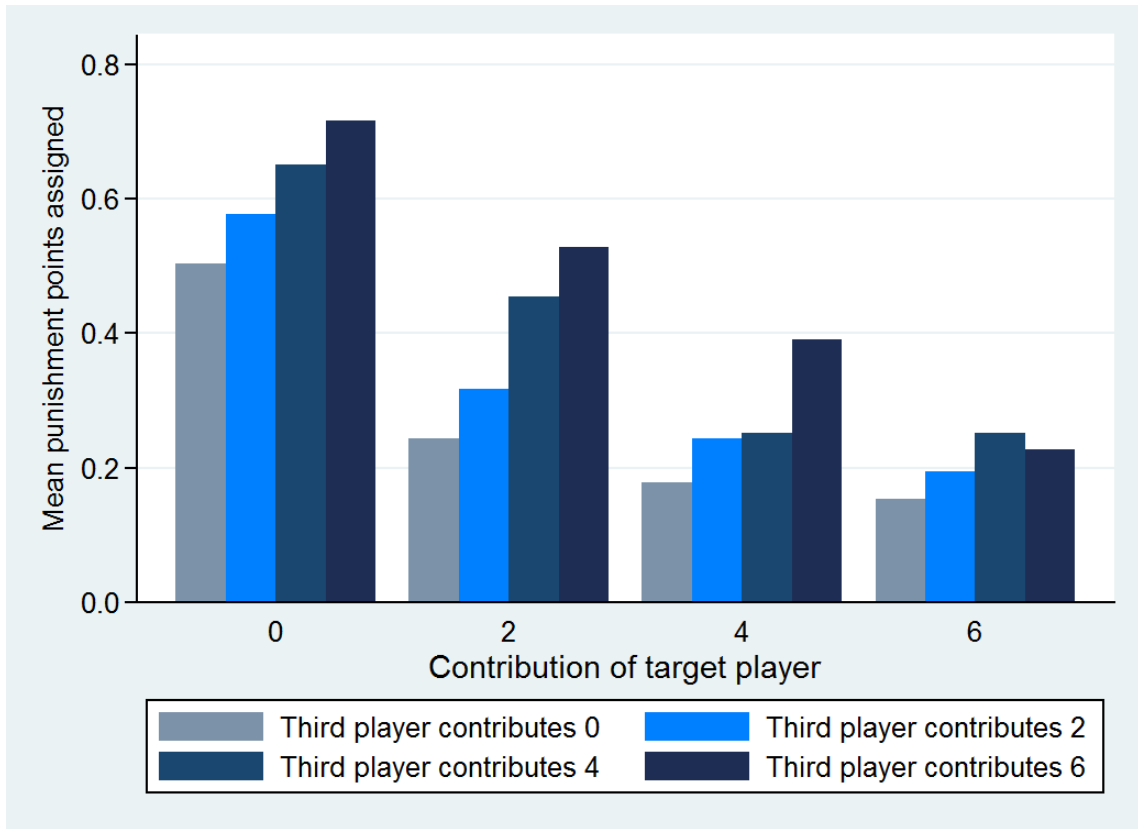


Figure 6. Mean punishment points assigned, as a function of the contributions of the punisher, target player, and third player. The data contains 40 subjects who contribute 0 in the game with punishment, 28 who contribute 2, 31 who contribute 4, and 24 who contribute 6.

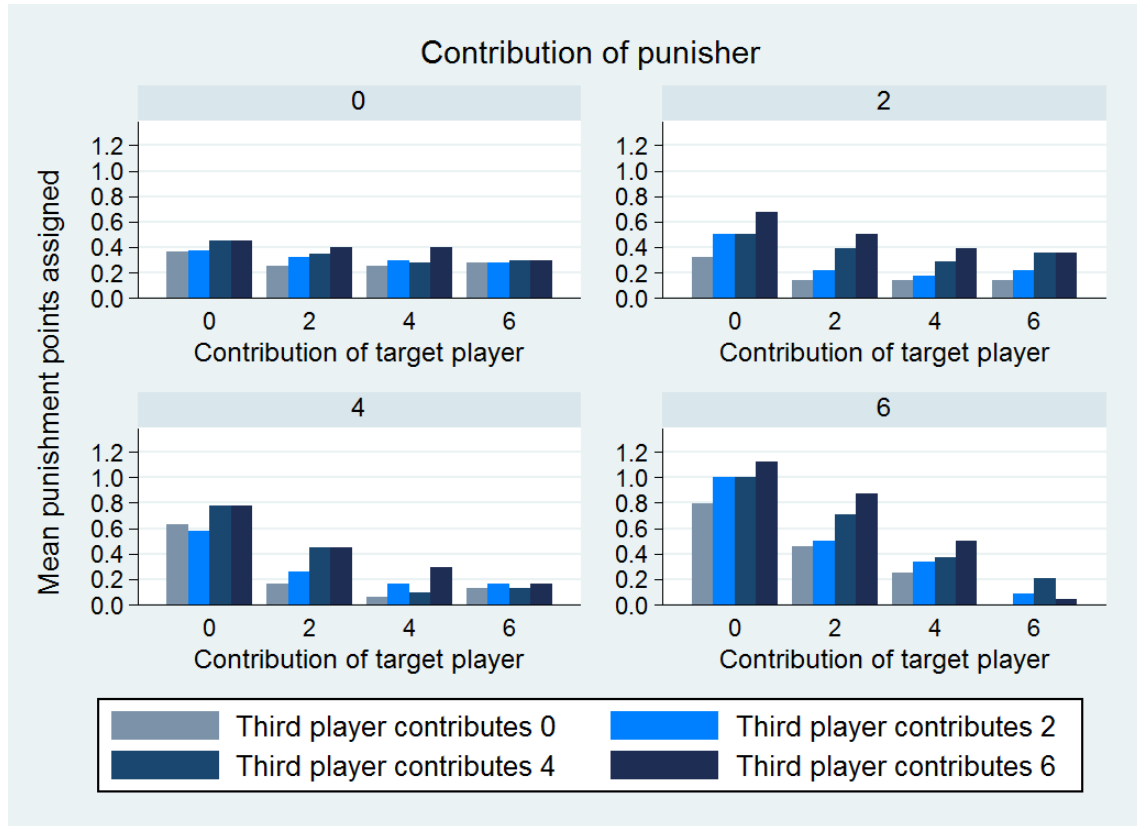


Table 1. Tests of two sources of selfish bias in conditional contributions.

	<b>Strong conditional cooperators</b>	<b>Conditional cooperators</b>	<b>Conditional cooperators and others</b>
Mean of c(2,2)	1.818	1.610	1.419
Signed-rank p-value: c(2,2) = 2	0.257	0.033	0.000
Mean of c(4,4)	3.394	3.024	2.387
Signed-rank p-value: c(4,4) = 4	0.002	0.000	0.000
Mean of c(6,6)	5.273	4.829	3.323
Signed-rank p-value: c(6,6) = 6	0.001	0.000	0.000
Mean of c(0,4)	1.091	1.024	0.935
Mean of c(2,2)	1.818	1.610	1.419
Signed-rank p-value: c(0,4) = c(2,2)	0.005	0.005	0.003
Mean of c(0,6)	1.515	1.415	1.484
Mean of c(2,4)	2.485	2.195	1.935
Signed-rank p-value: c(0,6) = c(2,4)	0.001	0.002	0.011
Mean of c(2,6)	2.727	2.488	2.032
Mean of c(4,4)	3.394	3.024	2.387
Signed-rank p-value: c(2,6) = c(4,4)	0.004	0.007	0.024
Number of subjects	33	41	62

Table 2: Random effects interval regressions of conditional contributions.

	Strong conditional cooperators			Conditional cooperators			Conditional cooperators and others		
	Coeff	SE	p	Coeff	SE	p	Coeff	SE	p
Mean of the other two players' contributions	0.941	0.040	0.000	0.907	0.039	0.000	0.751	0.046	0.000
Difference between the others' contributions	-0.169	0.029	0.000	-0.148	0.028	0.000	-0.035	0.034	0.308
Order (equals one for P-N sessions)	0.171	0.280	0.541	0.580	0.358	0.105	0.364	0.352	0.301
Constant	-0.374	0.253	0.140	-0.766	0.286	0.007	-1.037	0.305	0.001
p-value: $\sigma_u = 0$		0.000			0.000			0.000	
rho		0.468			0.627			0.458	
Log likelihood		-223.384			-287.764			-560.825	
Wald chi-square		612.35			597.31			273.04	
Number of subjects		33			41			62	
Number of observations		330			410			620	
Left censored observations		99			143			280	
Interval observations		202			235			305	
Right censored observations		29			32			35	
Marginal effects	dy/dx	SE	p	dy/dx	SE	p	dy/dx	SE	p
Mean of the other two players' contributions	0.921	0.038	0.000	0.833	0.039	0.000	0.564	0.038	0.000
Difference between the others' contributions	-0.166	0.028	0.000	-0.136	0.026	0.000	-0.026	0.026	0.309
Marginal effects	dy/dx	SE	p	dy/dx	SE	p	dy/dx	SE	p
Minimum of the other players' contributions	0.626	0.033	0.000	0.552	0.032	0.000	0.308	0.031	0.000
Maximum of the other players' contributions	0.295	0.035	0.000	0.281	0.033	0.000	0.256	0.033	0.000
p-value: minimum = maximum		0.000			0.000			0.308	

Table 3. Cross-tabulation of unconditional contributions ( $c_u$ ) by conditional contribution types in the game without punishment.

	Free-riders	Conditional cooperators	Others	Total
$c_u = 0$	52	8	9	69
$c_u = 2$	5	15	8	28
$c_u = 4$	4	11	4	19
$c_u = 6$	0	7	0	7
Total	61	41	21	123
Mean of $c_u$	0.426	2.829	1.524	1.415

Table 4. Mean beliefs regarding others' (unconditional) contributions, by unconditional contribution and conditional contribution type, in the game without punishment.

	Belief that others will contribute:				Expected Contribution
	0	2	4	6	
$c_u = 0$	66.59	16.25	9.36	7.80	1.17
$c_u = 2$	38.54	35.32	18.21	7.93	1.91
$c_u = 4$	24.00	26.74	36.63	12.63	2.76
$c_u = 6$	23.29	28.57	19.86	28.29	3.06
Free-riders	64.07	16.70	11.30	7.93	1.26
Conditional cooperators	34.61	29.29	23.80	12.29	2.28
Others	46.00	28.48	15.52	10.00	1.79
All subjects	51.16	22.91	16.19	9.74	1.69



Table 5. Random effects interval regressions of punishment points assigned.

	All subjects			N-P treatment order			P-N treatment order		
	Coeff	SE	p	Coeff	SE	p	Coeff	SE	p
Contribution of the punisher	-0.062	0.070	0.372	-0.123	0.083	0.138	-0.094	0.110	0.391
Contribution of the third player	0.107	0.012	0.000	0.073	0.014	0.000	0.132	0.018	0.000
Absolute negative deviation of target from punisher	0.355	0.020	0.000	0.348	0.027	0.000	0.362	0.027	0.000
Positive deviation of target from punisher	-0.070	0.019	0.000	-0.008	0.019	0.678	-0.159	0.036	0.000
Order (equals one for P-N sessions)	0.570	0.299	0.056						
Constant	-1.917	0.245	0.000	-0.929	0.390	0.017	-1.315	0.563	0.020
p-value: $\sigma_u = 0$		0.000			0.000			0.000	
rho		0.903			0.937			0.872	
Log likelihood		-930.875			-330.929			-573.709	
Wald chi-square		390.95			190.40			225.76	
Number of subjects		123			60			63	
Number of observations		2460			1200			1260	
Left censored observations		1893			964			929	
Interval observations		465			206			259	
Right censored observations		102			30			72	
Marginal effects	dy/dx	SE	p	dy/dx	SE	p	dy/dx	SE	p
Contribution of the punisher	-0.016	0.018	0.364	-0.029	0.019	0.128	-0.028	0.032	0.387
Contribution of the third player	0.028	0.004	0.000	0.017	0.004	0.000	0.039	0.007	0.000
Absolute negative deviation of target from punisher	0.092	0.009	0.000	0.081	0.011	0.000	0.107	0.014	0.000
Positive deviation of target from punisher	-0.018	0.005	0.000	-0.002	0.005	0.678	-0.047	0.011	0.000
p-value: negative deviation + positive deviation = 0		0.000			0.000			0.000	

Table 6. Marginal effects from random effects interval regressions of punishment, where coefficients are allowed to vary with the contribution of the punisher.

Marginal effects	All subjects			N-P treatment order			P-N treatment order		
	dy/dx	SE	p	dy/dx	SE	p	dy/dx	SE	p
Contribution of the punisher	-0.025	0.019	0.173	-0.045	0.020	0.021	-0.041	0.034	0.229
c = 0: Contribution of the third player	0.022	0.007	0.001	0.018	0.006	0.001	0.026	0.016	0.112
c = 0: Positive deviation of target from punisher	-0.030	0.007	0.000	-0.014	0.005	0.009	-0.080	0.020	0.000
c = 2: Contribution of the third player	0.042	0.007	0.000	0.017	0.006	0.008	0.063	0.012	0.000
c = 2: Absolute negative deviation of target from punisher	0.085	0.018	0.000	0.084	0.021	0.000	0.082	0.028	0.003
c = 2: Positive deviation of target from punisher	-0.010	0.010	0.308	0.021	0.011	0.053	-0.041	0.017	0.013
c = 4: Contribution of the third player	0.020	0.005	0.000	0.021	0.006	0.001	0.020	0.008	0.018
c = 4: Absolute negative deviation of target from punisher	0.094	0.011	0.000	0.088	0.014	0.000	0.107	0.018	0.000
c = 4: Positive deviation of target from punisher	0.010	0.018	0.571	0.047	0.020	0.019	-0.023	0.031	0.449
c = 6: Contribution of the third player	0.028	0.006	0.000	0.007	0.007	0.321	0.042	0.010	0.000
c = 6: Absolute negative deviation of target from punisher	0.097	0.010	0.000	0.096	0.015	0.000	0.109	0.016	0.000
p-value: c = 2, negative deviation + positive deviation = 0		0.001			0.000			0.262	
p-value: c = 4, negative deviation + positive deviation = 0		0.000			0.000			0.031	
p-value: sigma_u = 0		0.000			0.000			0.000	
rho		0.905			0.942			0.884	
Log likelihood		-921.013			-318.049			-562.908	
Wald chi-square		408.00			212.16			242.70	
Number of subjects		123			60			63	
Number of observations		2460			1200			1260	
Left censored observations		1893			964			929	
Interval observations		465			206			259	
Right censored observations		102			30			72	

Table 7. Mean beliefs regarding punishment incurred in the game with punishment (numbers in parentheses represent numbers of observations).

	<b>Own contribution</b>				<b>All subjects</b>
	<b>c = 0</b>	<b>c = 2</b>	<b>c = 4</b>	<b>c = 6</b>	
Never punish	1.00 (30)	0.92 (12)	1.33 (12)	0.50 (8)	0.98 (62)
Punish	2.90 (10)	2.06 (16)	1.32 (19)	1.38 (16)	1.79 (61)
Rank-sum p-value: Punish vs. never	0.002	0.057	0.766	0.323	0.004
Punish, never antisocial		1.00 (4)	0.82 (11)	0.67 (12)	0.78 (27)
Punish, sometimes antisocial	2.90 (10)	2.42 (12)	2.00 (8)	3.50 (4)	2.59 (34)
Rank-sum p-value: Antisocial vs. never		0.152	0.037	0.006	0.000

## APPENDIX A: SUPPLEMENTARY FIGURES

Figure A1. Decision screen for the contribution table of the strategy method game without punishment.

**Please complete the CONTRIBUTION TABLE for each combination of the contributions of the other two players in your group.  
Your contribution must take one of the values 0, 2, 4, or 6.**

**Contribution Table.** Please enter your contribution:

If the contributions of the other two players are <b>0</b> and <b>0</b> :	
If the contributions of the other two players are <b>0</b> and <b>2</b> :	
If the contributions of the other two players are <b>0</b> and <b>4</b> :	
If the contributions of the other two players are <b>0</b> and <b>6</b> :	
If the contributions of the other two players are <b>2</b> and <b>2</b> :	
If the contributions of the other two players are <b>2</b> and <b>4</b> :	
If the contributions of the other two players are <b>2</b> and <b>6</b> :	
If the contributions of the other two players are <b>4</b> and <b>4</b> :	
If the contributions of the other two players are <b>4</b> and <b>6</b> :	
If the contributions of the other two players are <b>6</b> and <b>6</b> :	

OK

Help

Your earnings = (6 - Your contribution) + (0.5 × Sum of all three players' contributions)

Figure A2. Sample decision screen showing one of the ten cases in the punishment stage of the strategy method game with punishment.

**Please enter the number of deduction points, if any, that you would assign to each player in each of the ten cases.  
Deduction points will only actually be assigned for the case that corresponds to the actual contributions of the other two players.  
You can assign 0, 1, 2, or 3 deduction points to each player in each of the cases.**

CASE 2 of 10.	YOU	Player B	Player C
Contribution:	0	0	2
Earnings from stage one:	7	7	5
Deduction points:	-		

Back
Next

Help

If you assign deduction points to another player, each deduction point will reduce the earnings of that player by three earnings points.  
For each deduction point that you assign to another player, you will incur a cost of one earnings point.

Figure A3. Mean punishment points assigned, as a function of the deviation in the contribution of the target player from the mean contribution of the punisher and third player, by treatment order.

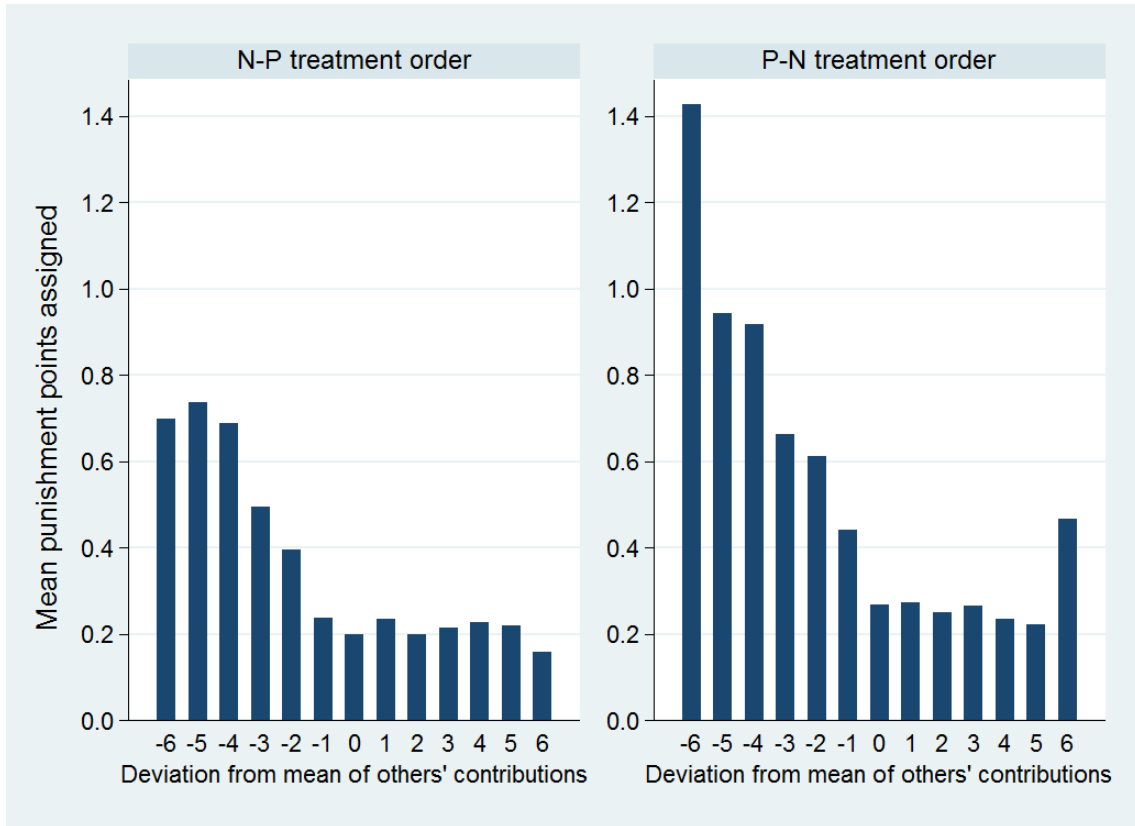


Figure A4. Mean punishment points assigned, as a function of the contributions of the target player and third player, by treatment order.

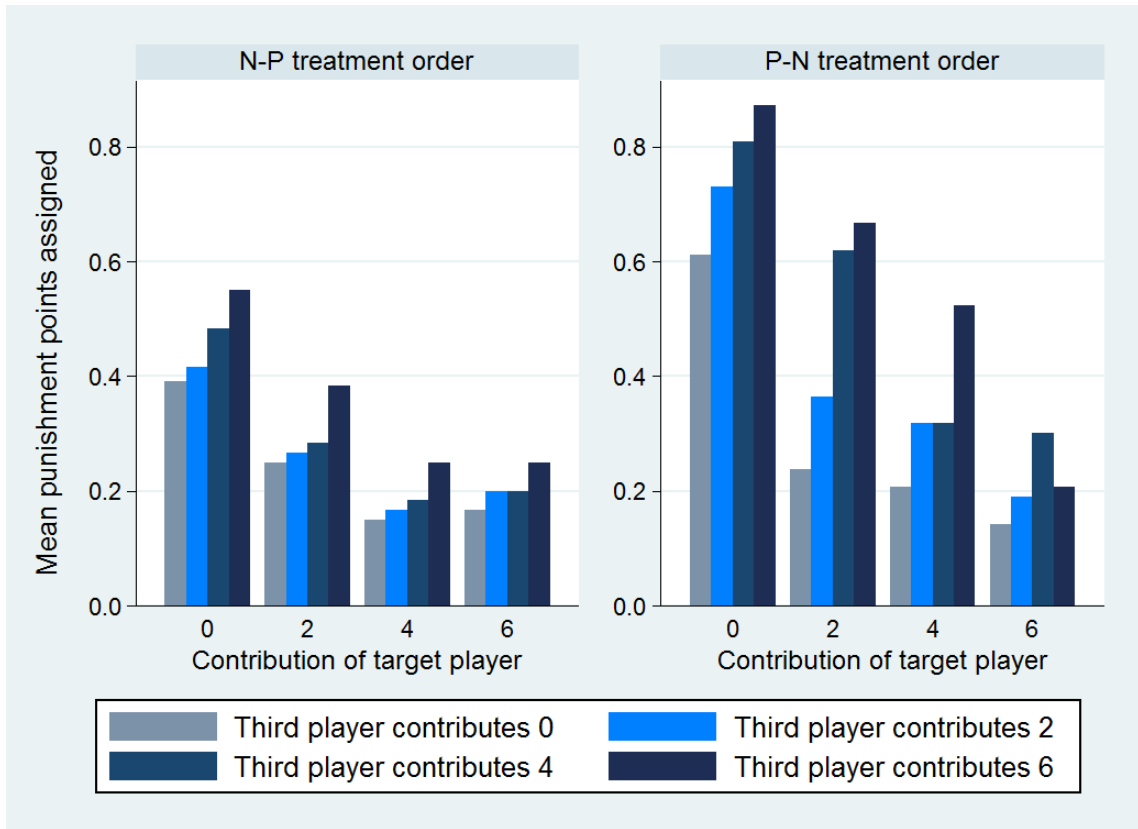


Figure A5a. Mean punishment points assigned, as a function of the contributions of the punisher, target player, and third player, N-P treatment order.

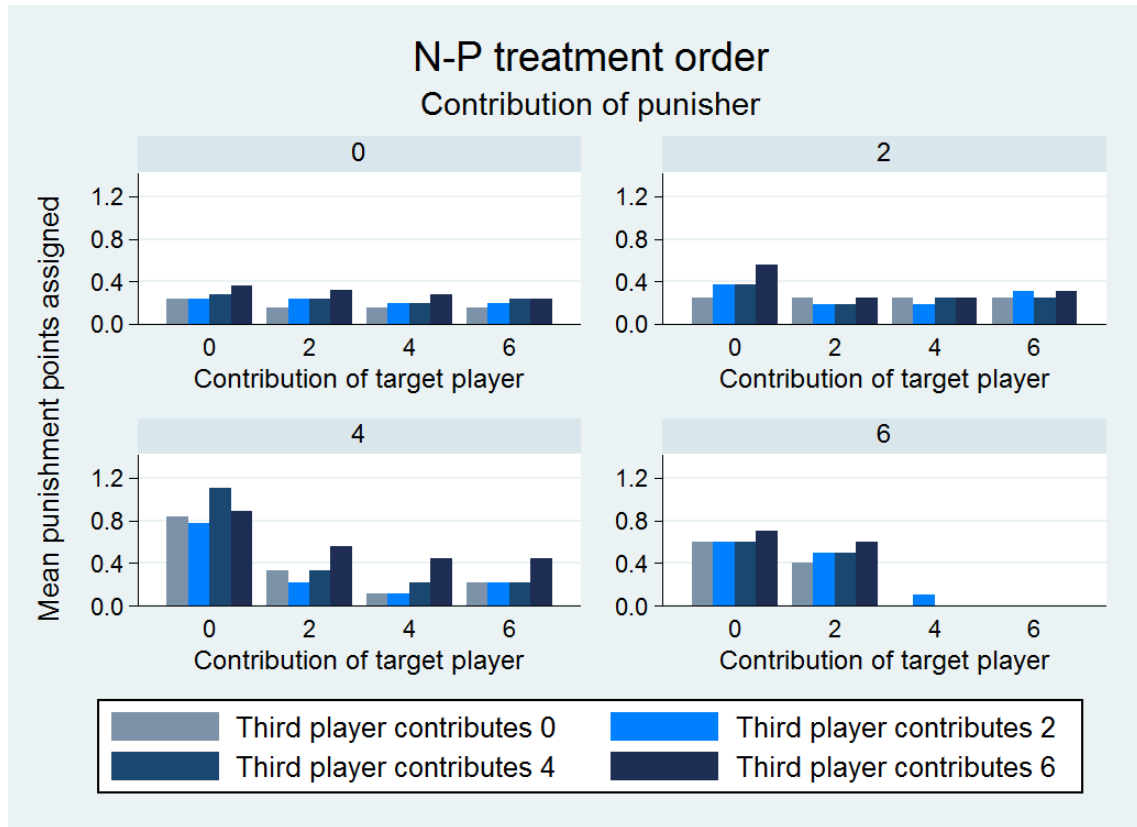
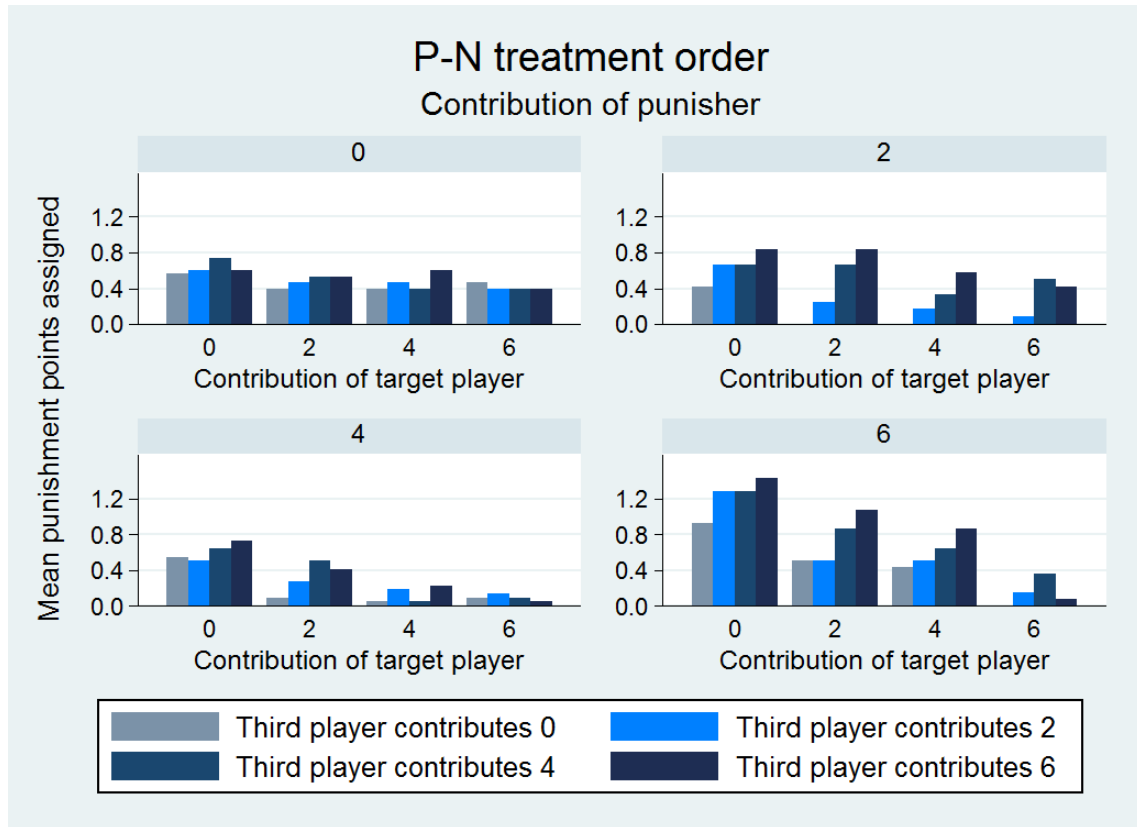


Figure A5b. Mean punishment points assigned, as a function of the contributions of the punisher, target player, and third player, P-N treatment order.





## APPENDIX B: INSTRUCTIONS FOR THE N-P TREATMENT ORDER

*Horizontal rules denote the positions of page breaks in the original instructions.  
Control questions were displayed on screen.*

### GENERAL INFORMATION

Welcome to today's session. In this session we will conduct two experiments on economic decision making. These experiments are simple, and if you read the instructions carefully and make good decisions, you may earn a considerable amount of money.

It is strictly prohibited to communicate with the other participants. If you violate this rule, you will be dismissed from the lab and forfeit all earnings. If you have any questions please raise your hand, and an experimenter will assist you.

In each of the two experiments, we will proceed through the following steps:

- Firstly, you will be given information about the decision situation for the experiment.
- You will then be asked to answer some questions at your computer to check that you fully understand the decision situation.
- Next, you will be given further instructions on how to use the computer screens to enter your decisions.
- Finally, you will enter your decisions into the computer.

At the end of the second experiment you will be given information about the results of your decisions in both experiments. After this, you will be asked to complete a questionnaire.

Throughout the session, your earnings will be calculated in "points". At the end of the session, the total number of points you have earned will be converted into Australian Dollars and paid to you in cash. You will be paid for your decisions in both experiments.

**The conversion rate will be 1 point = 1.5 Australian Dollars.**

At the beginning of the session, you will have a starting balance of three points. This is in addition to the other points that you earn throughout the session.

---

### FIRST EXPERIMENT: DECISION SITUATION

In this experiment, you will be in a group of three players, consisting of yourself and two others. All decisions will be made anonymously, and you will never learn the identity of the other two players in your group.

At the beginning of the experiment you will be given six points. (This is in addition to your starting balance of three points.) You have to decide how many of these six points you want to contribute to a project, and how many to retain for yourself.

**You can choose to contribute 0, 2, 4, or 6 points to the project.**

Each point that you do not contribute to the project will be automatically retained for yourself. The other two players will face the same decision situation.

### **Your earnings from this decision**

Your earnings will depend on both your own decision, and the decisions of the other two players in your group. These earnings consist of two parts:

1. Your earnings from the points you retain for yourself:

$$\text{Your earnings from points you retain for yourself} = 6 - \text{Your contribution.}$$

No-one other than you earns anything from the points you retain for yourself.

2. Your earnings from the project. This is calculated as:

$$\text{Your earnings from the project} = 0.5 \times \text{Sum of all three players' contributions.}$$

All three players receive the same earnings from the project. For each point that anyone contributes to the project, the total earnings of the group therefore increase by 1.5 points.

Your contribution to the project thus increases the earnings of the other two players. At the same time, you also receive earnings as a result of the other players' contributions.

Your total earnings are the sum of the points you retained and your earnings from the project:

$$\text{Your earnings} = (6 - \text{Your contribution}) + (0.5 \times \text{Sum of all three players' contributions})$$

The earnings of the other two players are calculated in the same manner.

---

Please answer the following questions. These serve to check your understanding of the decision situation and earnings calculations. When everyone has completed all the questions correctly, we will explain how the experiment itself will take place.

If you have any questions, please raise your hand.

1. Suppose that no-one contributes anything to the project.
  - a. What are your earnings in points?
  - b. What are the earnings of each of the other two players?
2. Suppose that all three players each contribute 6 points to the project.
  - a. What are your earnings in points?
  - b. What are the earnings of each of the other two players?
3. Suppose that you contribute 0 points, and the other two players each contribute 6 points.
  - a. What are your earnings in points?
  - b. What are the earnings of each of the other two players?

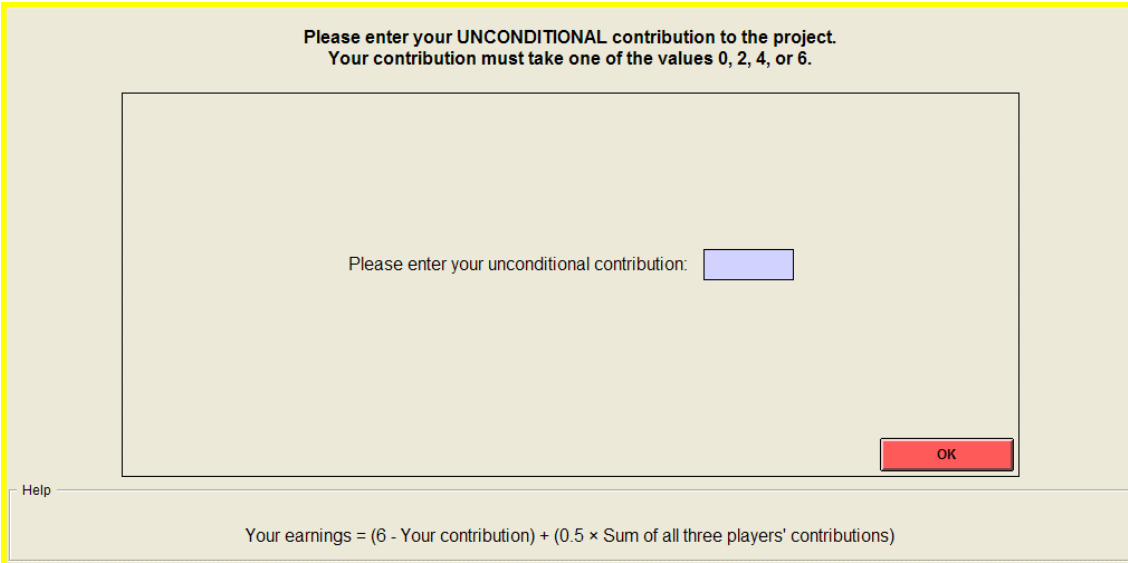
4. Suppose that the other two players contribute a total of 8 points to the project.
    - a. What are your earnings if you contribute 2 points?
    - b. What are your earnings if you contribute 4 points?
  5. Suppose that you contribute 4 points to the project.
    - a. What are your earnings if the other two players contribute a total of 4 points?
    - b. What are your earnings if the other two players contribute a total of 8 points?
- 

## FIRST EXPERIMENT: PROCEDURES

This experiment will take place only once. You will complete three tasks: an “unconditional” contribution, a “contribution table”, and your prediction of the contributions of the other participants in the laboratory.

### Unconditional contribution

In the “unconditional” contribution decision, you simply indicate how many points you want to contribute to the project. You can contribute 0, 2, 4, or 6 points. You enter your decision by typing one of these numbers in the input field on your screen:



The screenshot shows a web-based interface for an experiment. At the top, it says: "Please enter your UNCONDITIONAL contribution to the project. Your contribution must take one of the values 0, 2, 4, or 6." Below this is a large rectangular area containing the text "Please enter your unconditional contribution:" followed by a light blue input field. In the bottom right corner of this area is a red button labeled "OK". At the bottom left of the screen is a "Help" link. At the bottom center is the formula: "Your earnings = (6 - Your contribution) + (0.5 × Sum of all three players' contributions)".

After you enter the amount you want to contribute, you must click the OK button. As long as you have not clicked OK, you can still change your decision. After you have clicked OK, your decision can no longer be revised.

---

## Contribution table

In the “contribution table”, you indicate how many points you want to contribute to the project *for every possible combination of the contributions of the other two players* in your group. There are ten possible combinations, as you can see from the decision screen:

Please complete the CONTRIBUTION TABLE for each combination of the contributions of the other two players in your group.  
Your contribution must take one of the values 0, 2, 4, or 6.

**Contribution Table.** Please enter your contribution:

If the contributions of the other two players are 0 and 0 :	<input type="text"/>
If the contributions of the other two players are 0 and 2 :	<input type="text"/>
If the contributions of the other two players are 0 and 4 :	<input type="text"/>
If the contributions of the other two players are 0 and 6 :	<input type="text"/>
If the contributions of the other two players are 2 and 2 :	<input type="text"/>
If the contributions of the other two players are 2 and 4 :	<input type="text"/>
If the contributions of the other two players are 2 and 6 :	<input type="text"/>
If the contributions of the other two players are 4 and 4 :	<input type="text"/>
If the contributions of the other two players are 4 and 6 :	<input type="text"/>
If the contributions of the other two players are 6 and 6 :	<input type="text"/>

OK

Help

Your earnings = (6 - Your contribution) + (0.5 × Sum of all three players' contributions)

For each of the ten cases, you can choose to contribute 0, 2, 4, or 6 points to the project. You must make an entry in each of the ten input boxes. Once you have finished, please click OK.

Afterwards, the computer will randomly determine whether your unconditional contribution or your contribution table will be used to decide your earnings:

- For two of the three players in your group, the unconditional contribution will be used to decide that player’s contribution to the project.
- For the third player, the contribution table will be used. The computer will first look up the unconditional contributions of the first two players. It will then choose the appropriate contribution from the third player’s contribution table.

Your earnings will then be computed in the manner that was explained earlier:

$$\text{Your earnings} = (6 - \text{Your contribution}) + (0.5 \times \text{Sum of all three players' contributions})$$

Since you do not know, at the time you make your decisions, whether your unconditional contribution or your contribution table will be used to decide your earnings, you should treat both sets of decisions as if they would count for your earnings.

## Prediction of the other participants' contributions

You can earn additional points by predicting the *unconditional* contributions of the other participants in the laboratory. In particular, you will be asked in how many cases out of 100 you think the other participants contributed 0, 2, 4, and 6 points:

**Please enter your PREDICTION of the unconditional contributions of the other participants in the laboratory.**  
For each question you should enter a number between 0 and 100. The four numbers must add up exactly to 100.

In how many cases out of 100 do you think the other participants contributed **0** points?

In how many cases out of 100 do you think the other participants contributed **2** points?

In how many cases out of 100 do you think the other participants contributed **4** points?

In how many cases out of 100 do you think the other participants contributed **6** points?

For each of the four contribution levels you should enter a number between zero and 100. These numbers must add up exactly to 100. Afterwards, the computer will compare your predictions to the actual unconditional contributions of the other participants in the laboratory.

You can earn up to two extra earnings points for your predictions. *The closer your predictions are to the actual percentage of participants who chose each contribution level, the more you earn.* You cannot lose points from making predictions; it is only possible to earn more points.

The formula that determines your earnings from your predictions is as follows:

$$\text{Earnings from predictions} = 2 - \left(\frac{A-a}{100}\right)^2 - \left(\frac{B-b}{100}\right)^2 - \left(\frac{C-c}{100}\right)^2 - \left(\frac{D-d}{100}\right)^2$$

where:

$A$ = Percentage of the other participants who contribute 0,	$a$ = Your prediction of $A$ ,
$B$ = Percentage of the other participants who contribute 2,	$b$ = Your prediction of $B$ ,
$C$ = Percentage of the other participants who contribute 4,	$c$ = Your prediction of $C$ ,
$D$ = Percentage of the other participants who contribute 6,	$d$ = Your prediction of $D$ .

---

## SECOND EXPERIMENT: DECISION SITUATION

We will now conduct a new experiment, in which there are some changes. You will complete this second experiment before you learn the results from the first experiment. After this, you will learn the results from both experiments and there will be no further experiments.

In this new experiment, you will be in a new group of three players, consisting of yourself and two others. Your new group will not include either of the players you were grouped with in the first experiment. Again, you will never learn the identity of the other two players.

The new experiment consists of two stages.

### **Stage one**

Stage one is identical to the *unconditional* contribution in the first experiment. At the beginning of this stage you will again be given six points. You have to decide how many of these six points you want to contribute to a project, and how many to retain for yourself.

**You can choose to contribute 0, 2, 4, or 6 points to the project.**

Each point that you do not contribute to the project will be automatically retained for yourself. You will only make an unconditional contribution decision – there is no contribution table.

Your earnings from stage one will be computed in the same manner as in the first experiment:

$$\text{Earnings from stage one} = (6 - \text{Your contribution}) + (0.5 \times \text{Sum of all three players' contributions})$$

### **Stage two**

Stage two is new to this experiment. In this stage you can assign deduction points to reduce the earnings of one or both of the other players, or you can leave their earnings unchanged. The other players can also assign deduction points to reduce your earnings if they so wish.

**You can assign 0, 1, 2, or 3 deduction points to each of the other two players.**

If you assign deduction points to another player, *each deduction point will reduce the earnings of that player by three earnings points*. If you do not assign any deduction points to a player, then that player's earnings will be unchanged.

For each deduction point that you assign to another player, you will incur a cost of one earnings point. If you do not assign any deduction points, you will not incur any costs.

---

### **Your earnings from these decisions**

Your final earnings in this experiment will depend on your earnings from stage one, the number of deduction points you received from the other two players in your group, and the number of deduction points you assigned to them.

In particular, the computer will first take the number of deduction points (if any) you received from the other two players, and multiply this by three.

- If this amount (three times the number of deduction points you received) is no greater than your earnings from stage one, then your earnings will be reduced by this amount.

- Otherwise, if this amount is greater than your earnings from stage one, your earnings will be reduced to zero. Notice that this means that the deduction points you receive from the other players cannot cause you to suffer a loss.

After it has done this, the computer will then deduct the cost of any deduction points you assigned to the other two players. Notice that this means that you must always incur the cost of any deduction points you assign to the other players, even if this causes you to suffer a loss.

If your final earnings are negative, the loss will be taken out of the starting balance that you were given at the beginning of the session. Notice, however, that it is always possible to avoid such a loss for certain through your own decisions.

The earnings of the other two players will be calculated in the same manner.

---

Please answer the following questions. These serve to check your understanding of the decision situation and earnings calculations. When everyone has completed all the questions correctly, we will explain how the experiment itself will take place.

If you have any questions, please raise your hand.

6. Suppose that you assign 3 deduction points to the second player and 0 deduction points to the third player.
    - a. What cost do you incur to assign these deduction points?
    - b. By how much will the earnings of the second player be reduced?
    - c. By how much will the earnings of the third player be reduced?
  7. By how much will your earnings be reduced:
    - a. If you receive a total of 1 deduction points from the other players?
    - b. If you receive a total of 2 deduction points from the other players?
- 

## **SECOND EXPERIMENT: PROCEDURES**

This experiment takes place only once. You will complete four tasks: a contribution in stage one, a set of ten “cases” to assign deduction points in stage two, your prediction of the other participants’ contributions, and your prediction of the number of deduction points you receive.

### **Stage one contribution**

In the contribution decision in stage one, you indicate how many points you want to contribute to the project. You can contribute 0, 2, 4, or 6 points.

In this experiment *you only make an unconditional contribution decision – there is no contribution table*. This means that the amount you enter in the contribution screen will for certain be your contribution to the project.

## Stage two deduction cases

In the second task, you indicate how many deduction points, if any, you want to assign to each of the other two players.

When you assign deduction points, you will not know the actual contributions of the other two players. Instead, you will assign deduction points for ten cases, corresponding to *all ten possible combinations of the other two players' contributions*.

Of the two other players in your group, we refer to the one who contributes less to the project as “Player B”, and the one who contributes more as “Player C”. Then the ten cases are:

	Player B's contribution	Player C's contribution
Case 1	0	0
Case 2	0	2
Case 3	0	4
Case 4	0	6
Case 5	2	2
Case 6	2	4
Case 7	2	6
Case 8	4	4
Case 9	4	6
Case 10	6	6

You can assign up to three deduction points to each player in each of the ten cases. These deduction points will only actually be allocated for one of the ten cases. This is the case that corresponds to the actual contributions of the other two players.

For each case, you will complete a decision screen similar to the one shown for Case 1:

Please enter the number of deduction points, if any, that you would assign to each player in each of the ten cases. Deduction points will only actually be assigned for the case that corresponds to the actual contributions of the other two players. You can assign 0, 1, 2, or 3 deduction points to each player in each of the cases.

<b>CASE 1 of 10.</b>	<b>YOU</b>	<b>Player B</b>	<b>Player C</b>
Contribution:	...	0	0
Earnings from stage one:	...	...	...
Deduction points:	--	<input type="text"/>	<input type="text"/>

Help  
If you assign deduction points to another player, each deduction point will reduce the earnings of that player by three earnings points. For each deduction point that you assign to another player, you will incur a cost of one earnings point.



In each of the ten cases, your own contribution is always the amount you chose in stage one. In the actual experiment, each decision screen will show your own contribution in the first column, and you will also be able to see the earnings from stage one for each of the players.

For each of the cases, you must decide how many deduction points, if any, you want to assign to Player B and Player C. You must enter a number for each of the players. If you do not wish to reduce the earnings of a player, then you must enter “0”.

You can use the “Next” and “Back” buttons to move between the ten cases. The “OK” button appears after you have filled in all ten cases. As long as you have not clicked “OK”, you can change any of your decisions. After you click “OK”, your decisions can no longer be revised.

Afterwards, the computer will look up the stage one contributions of the other two players in your group. From this, it will determine which of the ten cases is the relevant one and use your decisions from that case to assign deduction points to the other two players.

The computer will also do the same for the two other players in your group. In this way it will determine how many deduction points, if any, you receive from each of the other players. Your earnings will then be computed in the manner that was explained previously.

### **Prediction of the other participants’ contributions**

You can again earn additional points by predicting the contributions of the other participants in the laboratory. As before, you will be asked in how many cases out of 100 you think the other participants contributed 0, 2, 4, and 6 points in the second experiment.

For each of the four contribution levels you should enter a number between zero and 100. These numbers must add up exactly to 100. Afterwards, the computer will compare your predictions to the actual contributions of the other participants in the laboratory.

You can earn up to two extra earnings points for your predictions. *The closer your predictions are to the actual percentage of participants who chose each contribution level, the more you earn.* You cannot lose points from making predictions; it is only possible to earn more points.

The formula that determines your earnings from your predictions is the same as before:

$$Earnings\ from\ predictions = 2 - \left(\frac{A-a}{100}\right)^2 - \left(\frac{B-b}{100}\right)^2 - \left(\frac{C-c}{100}\right)^2 - \left(\frac{D-d}{100}\right)^2$$

where:

$A$ = Percentage of the other participants who contribute 0,	$a$ = Your prediction of $A$ ,
$B$ = Percentage of the other participants who contribute 2,	$b$ = Your prediction of $B$ ,
$C$ = Percentage of the other participants who contribute 4,	$c$ = Your prediction of $C$ ,
$D$ = Percentage of the other participants who contribute 6,	$d$ = Your prediction of $D$ .

### **Prediction of deduction points received**

Finally, you will be asked to predict the total number of deduction points assigned to you by the other two players in your group. This must be a number between zero and six. Afterwards, the computer will compare your prediction to the actual number of deduction points you received.

You can earn up to one extra earnings point for your prediction. *The closer your prediction is to the total number of deduction points the other players assigned to you, the more you earn.* You cannot lose points from making predictions; it is only possible to earn more points.

The formula that determines your earnings from your prediction is as follows:

$$\text{Earnings from prediction} = 1 - \left( \frac{X - x}{6} \right)^2$$

where  $X$  is the actual number of deduction points you received, and  $x$  is your prediction of  $X$ .