UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

WORKING PAPERS IN ECONOMICS

No 451

Punishment Cannot Sustain Cooperation in a Public Good Game with Free-Rider Anonymity

Amrish Patel, Edward Cartwright and Mark van Vugt

May 2010

ISSN 1403-2473 (print)
ISSN 1403-2465 (online)

Department of Economics
School of Business, Economics and Law at University of Gothenburg
Vasagatan 1, PO Box 640, SE 405 30 Göteborg, Sweden
+46 31 786 0000, +46 31 786 1326 (fax)
www.handels.gu.se info@handels.gu.se

Punishment Cannot Sustain Cooperation in a Public Good Game with Free-Rider Anonymity

Amrish Patel
Department of Economics,
University of Gothenburg,
Box 640, SE 405 30,
Gothenburg,
Sweden.
Amrish.Patel@economics.gu.se

Edward Cartwright
Department of Economics,
Keynes College,
University of Kent,
Canterbury,
Kent. CT2 7NP. UK.
E.J.Cartwright@kent.ac.uk

Mark Van Vugt
Department of Work and Organizational Psychology,
VU University Amsterdam,
1081 BT Amsterdam,
The Netherlands.
M.van.Vugt@psy.vu.nl

May 18, 2010

Abstract

Individuals often have legitimate but publicly unobservable reasons for not partaking in cooperative social endeavours. This means others who lack legitimate reasons may then have the opportunity to behave uncooperatively, i.e. free-ride, and be indistinguishable from those with legitimate reasons. Free-riders have a degree of anonymity. In the context of a public good game we consider the effect of free-rider anonymity on the ability of voluntary punishment to sustain cooperative social norms. Despite only inducing a weak form of free-rider anonymity, punishment falls and cannot sustain cooperation.

Keywords: Anonymity, free-riding, public goods experiment, punishment

JEL codes: D82, C92, H41.

We gratefully acknowledge financial assistance from the British Academy through the BA Small Research Grant "Anonymous free-riding in collective action problems" (Grant number: SG50337) and helpful comments from Simon Gächter, Andreea Mitrut and Joel Sobel.

1 Introduction

A seminal finding is that voluntary punishment can sustain positive contributions to public goods (Fehr and Gächter 2000). Despite the cost of punishing, experimental subjects often punish those that under-contribute in order to encourage higher contributions (Carpenter et al. 2009) and express negative emotions towards fairness norm violations (Carpenter et al. 2004; Falk et al 2005; Hopfensitz and Reuben 2009; Casari and Luini 2010). This means that contributions are increased and the familiar free-rider problem is mitigated. The significance of this result cannot be understated, especially given suggestions that punishment provides an explanation for the evolutionary origins of cooperation in society (Fehr and Gächter 2002; Fehr and Fischbacher 2004).

Our objective in this paper is to question how much the success of punishment in sustaining cooperative social norms relies on the ability to *observe free-riders*. In many contexts, where contributions are publicized, we get to know who made a positive (or large) contribution but do not get to know who contributed zero (or little). We call this *free-rider anonymity.* One could infer that anyone whose contribution is not observed must have contributed zero. It seems crucial, however, to distinguish between people who *chose* to contribute zero and those who contribute zero through ignorance or *lack of opportunity.* For example, suppose that a head of department emails staff on Wednesday morning asking for volunteers to meet, say, prospective students and their parents on Wednesday afternoon. The head of department will see who turned up and who did not. What he will not know is *why* people did not attend. Was it because they did not want to go, or did they have an unavoidable meeting or engagement elsewhere?

Legitimate and publicly unobservable reasons to not contribute could lessen the possible benefits of punishment for two reasons. First, we know, from other contexts, that some individuals will 'hide' given the chance (Dana et al 2006; Broberg et al 2007; Lazear et al 2010) and free-rider anonymity gives that chance. Second, we know that punishment and fairness depend crucially on beliefs about intentions (Rabin 1993; Charness and Dufwenberg 2006; Falk et al 2008) and free-anonymity makes intentions unclear. To illustrate, consider the head of department dealing with those who did not turn up but not being able to distinguish those with and without a legitimate reason for not showing up. To punish all those who did not turn up might annoy those who had a legitimate reason to not go. But, to not punish anyone let's off the free-riders.

Can punishment still be effective in such a setting? We find that it is not. This seems a crucial insight because an inability to observe free-riders is not only relevant for the head of department on a Wednesday afternoon but common in many situations (Hugh-Jones and Reinstein 2009). The degree of free-rider anonymity may, therefore, be crucial in understanding social norms.

To briefly explain how we derive this conclusion, we have subjects play a public good game in which there are *six* group members of which only *four* have the opportunity to contribute. Two players thus have a legitimate reason for contributing zero. We find that in conditions where it is possible (but not necessarily obvious) to work out which group members could have contributed but did not, contributions increase when punishment is permitted. This is consistent with the standard finding. By contrast, when it is not possible to work out whether a group member deliberately contributed zero or not we find that contributions decline despite punishment opportunities.

In relating our work to the literature we first recognize that punishment is known to be no panacea. Recent research efforts have, therefore, rightly turned to identifying the limits of when punishment can sustain cooperation. The effectiveness of punishment has been shown to depend on the cost of punishment (Anderson and Putterman 2006; Carpenter 2007a; Nikiforakis and Normann 2008; Egas and Riedl 2008), societal norms and values (Cinyabuguma et al 2006; Herrmann et al 2008; Gächter and Herrmann 2009; Gächter and Herrmann 2010), and the possibility of counter-punishment (Denant-Boemont et al 2007; Nikiforakis 2008).[1] Our work clearly complements this literature by providing another piece to the puzzle of when punishment can sustain cooperation. We think that it is a fundamental piece given the possible importance of free-rider anonymity.

In almost all research on punishment in public good games subjects observe the contributions of all group members. Free-rider anonymity is one of several different forms of *imperfect monitoring*, where not all contributions are observable, that exist in the real world. An emerging literature has begun to explore the consequences of imperfect monitoring; we have learnt that punishment is less effective the smaller the set of individuals that can be monitored and punished (Carpenter 2007b), if endowments are imperfectly known (Bornstein and Weisel 2010), and if observed contributions are subject to noise (Grechenig et al. 2010). [2] [3] Our work differs from these studies in two fundamental respects. Free-rider anonymity means that the ability to observe cooperativeness is, (1) correlated with the level of cooperativeness (or lack thereof) and (2) endogenously determined by the contributor[4]. Neither of these properties are present in the previous studies, but they do seem natural and important to consider. Indeed, because the degree of monitoring is endogenous we cannot say whether it is higher or lower in our study compared to others; if all subjects make a positive contribution there is complete information, but if all subjects contribute zero, information is very much imperfect.

The rest of the paper is structured as follows. We describe our simple modification to the standard public goods game in Section 2. In Section 3 we state

[1] The efficacy of punishment is also reduced if there are privileged group members, those who have an incentive to provide the public good even if others do not (Reuben and Riedl 2009), and if payoff rather than contribution feedback is received (Nikiforakis 2010).

[2] A closely related idea is that of *imperfect punishment* where there is uncertainty over whether punishment will actually be carried out. Sousa (2010) demonstrates cooperation cannot be sustained with such an institution.

[3] Imperfect monitoring is not solely of interest in the punishment and public goods literature, it has also been studied in common resource management (Tarui et al. 2008) for example.

[4] Permitting endogenous choice of other aspects of institutions has been found to have significant effects on cooperation (Gürerk et al. 2006; Sutter et al. 2010).

our hypotheses on how we expect free-rider anonymity to affect behavior. Our results are found in Section 4, followed by our conclusion in Section 5.

2 Experimental design

We modify the standard protocol for linear public goods experiments slightly (e.g. Fehr and Gächter 2000) to test for the effects of free-rider anonymity.

Each subject is randomly assigned to a group of six participants with whom they play a linear public goods game for 20 periods. *Partners* matching and fixed IDs are intentionally used to activate reputation concerns and make cooperation easier. The first 10 periods involve only a *Contribution Stage*, the last 10 periods involve a *Contribution Stage* followed by a *Punishment Stage*[5]. These are as follows:

Contribution Stage: At the start of each period, each of the 6 subjects is given an endowment of 20 points (1 point = £0.02) and 4 of the 6 subjects are randomly selected to *take part* in the period. We shall refer to the 4 selected as *included* subjects and the other two as *excluded* subjects. Included subjects have the opportunity to split their endowment between a group account and a private account. Subjects' decisions are made simultaneously. Subject i's earnings in period t at the end of the contribution stage are given by

$$\pi_{it} := 20 - g_{it} + 0.4 \sum_{j \in N} g_{jt}$$

where g_{it} is subject i's contribution in period t and N is the set of subjects in i's group. In other words, each point a subject contributes to the group account costs him 0.6 points but will increase the earnings of all other group members (both included and excluded subjects) by 0.4 points. This is the standard linear public good game with the stark trade-off between self-interest (contributing zero) and group efficiency (contributing the endowment).

Punishment Stage: In periods 11-20, after the contribution stage subjects have the opportunity to punish each other. Both included and excluded subjects may punish and be punished. A subject can reduce another subject's earnings by 3 points at a cost of 1 point to themselves. They may punish as much or as little as they wish providing the cost of their punishment is less than π_{it}. Decisions are again made simultaneously. Once all punishment decisions have been made, subject i's total period earnings for period t are

$$\prod\nolimits_{it} := \pi_{it} - \sum_{j \in N} p_{it}^{j} - 3 \sum_{j \in N} p_{jt}^{i}$$

[5] Our design is not counter-balanced, that is, it is always the case that the first 10 periods do not have punishment stages and the final 10 periods do. A priori, there is little reason to suspect our main results would change if we were to reverse the order. Previous studies have not documented order-effects that qualitatively change punishment and contribution behavior (e.g. Fehr and Gächter 2000; Gächter and Herrmann 2010), however future work may want to test our results are robust to order-effects.

where p_{it}^j is the number of punishment points given by subject i to subject j in period t.

At the end of the 20 periods subjects are paid their cumulative earnings in cash[6] and a small show-up fee.

Other than the random selection of 4 out of 6 subjects taking part in each period, the game has been studied widely. To identify the effects of free-rider anonymity we use a 2×2 between-subject design. The two factors we vary are *Exclusion Information (EI)* and *Announcing Zeros (A0)*:

- **Exclusion Information (EI):** In an EI treatment subjects are told which subjects are included and excluded at the start of each period. In a treatment without exclusion information (NEI), this information is absent.

- **Announcing Zeros (A0):** In an A0 treatment all contributions (and the subjects who made them) will be listed at the start of the punishment stage (excluded subjects are always omitted from the list). In a treatment not announcing zeros (NA0) only non-zero contributions (and the subjects who made them) will be listed.

The 4 treatments are naturally coded: EI-A0; EI-NA0; NEI-A0; and NEI-NA0. We shall often refer to treatment NEI-NA0 as the one with *free-rider anonymity*.

EI-A0 is the *baseline* as this is very close to how public goods experiments are usually run. The only difference is that two subjects benefit from the public good despite having no opportunity to contribute. In EI-NA0, since it is known which 4 subjects are included, any of these 4 that do not appear on the list of contributors must have contributed zero. In NEI-A0 it is directly possible to see in the punishment stage who could have contributed but did not. Both EI-NA0 and NEI-A0 differ, therefore, from the baseline treatment in a, theoretically, trivial way. The NEI-NA0 treatment is fundamentally different as free-riders cannot be identified. There is *free-rider anonymity*. For example, if there are only two positive contributions made (and thus two IDs listed), this leaves 4 subjects, two of which contributed zero and the other two who were excluded. It is impossible to distinguish one from the other.[7]

The experiment was conducted at the University of Kent and our subjects were drawn from the student population. Mean earnings were £8.50. The table below states the number of groups in each treatment. A session consisted of three groups participating simultaneously[8].

<Table 1: Treatments>

[6] Including a one-off 25 points each subject would receive during the experiment if their earnings became negative. This was not used by any subject thus the 25 points were added at the end.

[7] Even this treatment is a weak form of anonymity, as we can calculate *how many* are free-riding, just not *who* is free-riding. One would expect any effects we find to be significantly larger in real-world settings where we cannot even deduce how many free-riders there are.

[8] Note that in one of the EI-A0 sessions there were only two groups due to a lack of subjects. Also note that technical difficulties at the start of period 18 meant we only have data up to period 17 for a different EI-A0 session.

3 Hypotheses

Like most linear public good games the subgame perfect Nash equilibrium is for all subjects to neither contribute nor punish. As explained in the introduction however, scores of studies have shown that subjects do both; we thus propose three more behavioral hypotheses.

Hypothesis 1: Contributions are the same across all treatments in periods 1-10. Contributions and punishment are the same across treatments excluding that with free-rider anonymity in periods 11-20.

The first part of Hypothesis 1 follows from the absence of procedural differences by treatment in periods 1-10. The second part follows from the observation that all contributions are publicized, albeit by different methods, in treatments other than that with free-rider anonymity.

Hypothesis 2: Free-riders are punished less with free-rider anonymity.

Assuming subjects wish to punish free-riders[9], does a free-rider deserve the same punishment in all treatments? Arguably in the free-rider anonymity treatment not only is an individual free-riding but they are also trying to "hide" their free-riding. Since we care about intentions (Falk et al. 2008) free-riders in NEI-NA0 may therefore deserve more punishment. However, we would argue this outcome is unlikely given that it is straightforward to punish free-riders in all treatments except NEI-NA0. With free-rider anonymity, suppose there is one free-rider and three potential candidates, how do you punish? There are several possibilities. First, you could punish all three severely. While individuals enjoy punishing (de Quervain et al 2004), this strategy involves punishing excluded subjects who did not have the opportunity to contribute, violating fairness norms (Rabin 1993) and causing resentment, discouraging future contributions from these individuals. Furthermore the strategy is very costly which may inhibit punishment (e.g. Anderson and Putterman 2006). This approach seems unlikely. A second strategy is to punish all three but less severely, trading-off the vigor of punishment against its cost. Yet another strategy is to maintain the severity but punish a randomly selected proper subset of the three. This also reduces the cost, but means free-riding could go unpunished. All three of these strategies involve potentially punishing subjects who did not have the opportunity to contribute, violating fairness and breeding resentment. Not punishing altogether is yet another option. While it avoids unjust punishment, free-riding does go unpunished. All except the first strategy imply Hypothesis 2.

Hypothesis 3: Contributions are lower and free-riding more common in periods 11-20 with free-rider anonymity.

[9] A non-trivial assumption, given anti-social punishment (e.g. Herrmann et al 2008).

Recall that in our free-rider anonymity treatment the degree of monitoring imperfection is endogenously determined. One could argue the opposite to Hypothesis 3 by asserting that individuals will recognize that free-riding hinders monitoring efforts and thus group cooperation. However, we would expect this to be quite unlikely given that individuals do act opportunistically if they can do so and get away with it (e.g. Dana et al 2006). Hypothesis 3 is based on a number of compelling and plausible effects. Research suggests individuals contribute more when they are identified, even if only by ID code, than when they are anonymous (Sell and Wilson 1991). This implies Hypothesis 3 once we note that NEI-NA0 is the only treatment where subjects have a degree of anonymity. Some of the punishment strategies discussed when justifying Hypothesis 2 involved excluded subjects being unintentionally punished when trying to punish actual free-riders. As mentioned, these individuals may feel resentment and thus reduce their contributions when they are next included. Another feature of some of the proposed punishment strategies is that some free-riders may go unpunished; it is well established that free-riders will not increase their contributions where this is true (e.g. Fehr and Gächter 2000). More generally when a punishment system is perceived as procedurally unfair research would suggest less cooperation (van Prooijen et al 2008). Finally, although those choosing to contribute zero cannot be punished easily, those that free-ride at a more moderate level can be, e.g. contributions of only 1 or 2. However, rather than subsequently increasing their contributions, one would expect these individuals to reduce their contributions all the way to zero in order to try and avoid being punished. All of these effects imply Hypothesis 3.

4 Results

In analyzing the data we first examine how contributions were affected by our treatments, then consider differences in punishment, finally we test whether the punishment differences explain the contribution differences.

Result 1: Punishment only fails to sustain contributions when there is free-rider anonymity.

Figure 1 compares mean contributions[10] in the free-rider anonymity treatment with the other treatments by period. Note that non-free-rider anonymity treatments are grouped in figures only for exposition purposes. Our statistical tests do not aggregate across treatments. Result 1 seems clear from Figure 1. Punishment can sustain contributions for all treatments except free-rider anonymity: final period contributions are a third lower than those of period one and almost 50% less than those of other treatments' final period.

<Figure 1: Contributions>

[10] We restrict the sample to only included subjects for all the contribution analysis that follows: excluded subjects do not make a contribution decision by definition.

Mean contributions without punishment, i.e. periods 1-10, (EI-A0 = 5.520; EI-NA0 = 5.821; NEI-A0= 7.142 and NEI-NA0= 5.552) are relatively similar across treatments. Treating each group's mean contribution over the relevant 10 periods as an independent observation, i.e. the number of observations is equal to the number of groups, a Kruskal-Wallis test suggest no significant differences in contributions across treatments ($\chi^2 = 2.487$, $p = 0.4776$). This is what we would expect. We also find, however, a similar result with punishment, i.e. periods 11-20, (EI-A0 = 6.156; EI-NA0 = 6.579; NEI-A0= 7.813 and NEI-NA0= 4.808). Mean contributions are lower with free-rider anonymity but we still find no significant differences in contributions across treatments (Kruskal-Wallis test[11]: $\chi^2 = 4.123$, $p = 0.2485$). It is however premature to conclude there are no treatment differences as we will see.

By comparing contributions within a treatment, without and with punishment, we can use more powerful matched-pairs tests. Again taking each group's mean contribution over the relevant 10 periods as an independent observation, Wilcoxon matched-pairs tests only detects a significant difference in the free-rider anonymity treatment (NEI-NA0: $z = 1.726$, $p = 0.0844$). This suggests cooperation is sustained by punishment in all treatments except free-rider anonymity, where it declines.

Returning to Figure 1, the above tests overlook important temporal differences in the punishment stage. With punishment, it seems that contributions are similar across treatments for the first 4 periods and then are lower in the free-rider anonymity treatment for the final 6 periods (we shall see shortly this can be traced to subjects initially trying to use punishment in this treatment and then stopping). Indeed a Kruskal-Wallis test indicates no significant differences across treatments for the first 4 periods ($\chi^2 = 3.373$, $p = 0.3376$). Using Wilcoxon Rank Sum tests to compare mean contributions in the final 6 periods across pairs of treatments we find that the only significant difference is between free-rider anonymity and the A0 treatments (EI-A0: $z = 1.908$, $p = 0.0564$; EI-NA0: $z = 1.405$, $p = 0.1601$; NEI-A0: $z = 1.780$, $p = 0.0752$), implying contributions are significantly lower in free-rider anonymity than these treatments.

For a more robust result we estimate a Tobit regression model for individual contributions left-censored at zero and right-censored at 20[12]. The explanatory variables consist of each of the treatment dummies interacted with dummies for the no punishment and punishment phases, thus separating out contributions in each treatment-phase. We use robust standard errors clustered by group to control for unobserved cross-sectional group heterogeneity[13] and include period-level fixed-effects to capture time trends.

[11] As we have done so far, all non-parametric statistical tests henceforth will take the average of the variable of interest over some number of periods per group as an independent observation. That is, the number of observations will equal the number of groups in the treatment.

[12] The results of all censored Tobit models we present are robust to estimating OLS models with the sample restricted to non-censored observations.

[13] The results of this model and others reported in this paper are robust to clustering standard errors by individual rather group and including individual-specific random-effects.

<Table 2: Determinants of contributions>

The estimates agree with the non-parametric tests: if we ignore the case of free-rider anonymity with punishment, we cannot reject the null that the remaining interaction coefficients are equal to zero ($F = 1.01$, $p = 0.4187$), thus contributions are the same across all these treatments, with and without punishment. Comparing coefficient estimates within treatment, we see contributions are only significantly different with punishment for the free-rider anonymity treatment ($F = 7.94$, $p = 0.001$)[14], punishment thus reduces contributions in this treatment. Overall then, the contribution data suggest punishment fails to sustain cooperation with free-rider anonymity.

For further evidence of the effect, consider the share of subjects who contribute nothing.

<Figure 2: Share of included subjects contributing zero>

We see the familiar pattern where treatments are broadly indistinguishable until period 14 (Kruskal-Wallis test: $\chi^2 = 3.561$, $p = 0.3129$), in the final 6 periods however the share of non-contributors is significantly different across treatments (Kruskal-Wallis test: $\chi^2 = 11.60$, $p = 0.009$) driven by the large number of non-contributors in the free-rider anonymity treatment. Almost 50% of all included subjects contribute nothing in the final period with free-rider anonymity, relative to a mere 15% in other treatments. All this is strong evidence for Hypotheses 1 and 3.

To understand why free-rider anonymity causes a collapse in cooperation we next consider punishment behavior.

Result 2: There is less punishment when there is free-rider anonymity, particularly of free-riders.

The figure below compares punishment points dispensed in the free-rider anonymity treatment relative to the others.

<Figure 3: Punishment dispensed>

We see that punishment starts at a similar level but soon diverges, with punishment in the free-rider anonymity treatment ending up less than half that in the other treatments. The mean level of punishment appears to be higher for the A0 treatments (EI-A0 = 3.489; EI-NA0 = 2.158; NEI-A0= 3.211 and NEI-NA0= 2.042), indeed a Kruskal-Wallis test confirms a significant difference across treatments ($\chi^2 = 7.886$, $p = 0.0484$). However, pairwise treatment comparisons indicate the only significant difference is between the baseline and

[14] It is somewhat surprising that punishment did not lead to significantly *higher* mean contributions in the baseline treatment, but merely non-declining contributions. That punishment stabilises contributions is a common finding, its level varies with subject pool (Herrmann et al 2008). Since our interest is primarily in the contrast between free-rider anonymity and the other treatments, this point is not critical.

free-rider anonymity (Wilcoxon Rank Sum test: $z = 2.216$, $p = 0.0267$), with subjects dispensing more punishment in the former than the latter.[15]

For further evidence, we estimate two Tobit regression models for the determinants of punishment dispensed, first for the entire dataset and then restricting it to included subjects only. In both models, the dependent variable is left-censored at zero and right-censored at π_{it}, robust standard errors are clustered by group to control for unobserved group heterogeneity and we include period-level fixed-effects to capture time trends. We use treatment dummies to capture any differences between treatments and also include some controls as explanatory variables. The average contribution and contribution variance in that period are included since one would expect less punishment if the group is making high contributions and more punishment if there is a large range of contributions. For the included subjects only model, we include the contribution the individual makes as an explanatory variable as one may expect the more he contributes the more he is likely to punish others[16]. The individual's contribution is not meaningful for the entire dataset model as excluded subjects do not have the opportunity to contribute, so we simply include a dummy to distinguish between excluded and included subjects. Finally since an individual must pay for their punishment points from their earnings in the contribution stage of a period, we also include their earnings in that period as an explanatory variable.

<Table 3: Determinants of punishment dispensed>

In line with the non-parametric tests, there are no significant differences in punishment dispensed across treatments excluding free-rider anonymity (All subjects: $F = 1.21$, $p = 0.2975$; Included only model: $F = 0.75$, $p = 0.4714$). The coefficient on the free-rider anonymity treatment dummy is large, negative and statistically significant, there is thus significantly less punishment in free-rider anonymity than the other treatments.

Not only is there less punishment in this treatment, there is also evidence that the punishment that there is less targeted at free-riders. For example, consider the punishment received by included subjects who contributed zero. These free-riders were both punished less in the free-rider anonymity treatment relative to the baseline (68% versus 93%; Wilcoxon Rank Sum test: $z = 2.333$, $p = 0.0197$) and also received fewer punishment points on average (1.6 versus 7.6 points; Wilcoxon Rank Sum test: $z = 3.128$, $p = 0.0018$). This results in the free-rider anonymity treatment being the only one where excluded subjects receive the same level of punishment as included subjects (Wilcoxon Rank Sum test: $z = 1.415$, $p = 0.1570$), in other treatments included subjects receive roughly double the punishment of excluded ones.[17]

[15] Similar results are found if we consider propensity to punish instead. Note also, there is no statistically significant difference in the punishment decisions of included and excluded subjects.

[16] Instead of this variable, we also estimated a model with the deviation of the punisher's contribution from the average of others' contributions and found the same results.

[17] It is surprising how often excluded subjects are punished in general (e.g. 59% of the time in the baseline treatment). We leave it to future research to determine whether motives such

More generally, this amounts to reduced punishment of individuals who contribute less than the mean of others' contributions[18] in the free-rider anonymity treatment as shown in the figure below.

<Figure 4: Punishment by contribution deviation>

It would be inappropriate to assume that the deviation of an excluded subject's contribution (zero by default) from that of others' would determine the punishment he receives[19], thus Figure 4 focuses on included subjects. Unlike other treatments where there is a significant negative correlation between the contribution deviation and punishment received, that in free-rider anonymity is positive and statistically insignificant (Kendall's Tau test: $\tau_b = 0.0543$, $p = 0.5888$) because of under-contributors going unpunished.

For a more robust picture of how different contributions are punished across treatment we estimate a Tobit model of punishment received restricting the sample to included subjects only for the reason we just mentioned. The dependent variable is left-censored at zero[20], as usual, we cluster robust standard errors by group and include period fixed-effects. We interact each of the treatment dummies with the absolute value of the negative and positive deviation of a subject's contribution from the average of other group members' contributions in the current period and with a dummy indicating whether they are free-riding. This allows us to distinguish pro-social, anti-social and free-rider punishment by treatment. We also include other group members' average contribution in that period as a control variable.

<Table 4: Determinants of punishment received>

There are two important inferences to be drawn from these estimates. The first is that in all treatments except free-rider anonymity, the more a subject's contribution is below the mean of others' contributions, the more they are punished. Indeed negative deviations are punished no differently from positive deviations in free-rider anonymity ($F = 1.41$, $p = 0.2439$). The second is that those that contribute nothing, free-riders, are punished significantly less than other contribution levels in free-rider anonymity[21]. All this evidence supports Hypotheses 1 and 2, free-riders are punished less with free-rider anonymity.

as inequity-aversion can explain this behavior.

[18] We obtain similar results if we consider the deviation of the punished subject's contribution from the punishing subject's contribution.

[19] Indeed the correlation is not statistically significant (Kendall's Tau test: $\tau_b = -0.0405$, $p = 0.7438$).

[20] We could also right-censor the dependent variable with the earnings of other group members in the contribution stage. However there is no need as this constraint is not binding: recall in Table 3 we saw that only one subject ever spent all their earnings on punishment.

[21] The significant coefficients for the intermediate treatments indicating antisocial punishment in NEI-A0 and less free-rider punishment in EI-NA0 are somewhat surprising. We expect the former is due to a few subjects in this treatment who were strong antisocial punishers and the latter due to the fact that there were very few free-riders in this treatment. Recall we only have 6 groups in each of these two treatments, these effects would probably disappear with more observations.

Finally, we now demonstrate that the adverse punishment profile uncovered under free-rider anonymity causes the observed cooperation collapse.

Result 3: Subjects contributing less than others do not increase contributions even if they are punished when there is free-rider anonymity.

To identify this effect we estimate a Tobit model for the determinants of a subject's change in contribution. For a clean test, the sample is restricted to only subjects included in periods t and $t+1$[22]. We left-censored the dependent variable at $(0 - contribution_{it})$ and right-censored it at $(20 - contribution_{it})$, with robust standard errors clustered by group and period-level fixed-effects. The model is estimated separately for individuals who contribute less than the mean of others' contributions, low contributors, and those who contribute more than the mean of others', high contributors, as one would expect these two groups to change their contributions in different directions following punishment. We interact each of the treatment dummies with the amount of punishment an individual receives in period t and the absolute deviation of their contribution from the mean of others' contributions, thereby allowing us to distinguish the effect of punishment from the effect of just contributing differently from others.

<Table 5: Determinants of the change in contribution>

For low contributors it seems that either subjects increase their contributions more the further below others' contributions theirs is (EI-A0) and punishment has no effect[23], or they do not automatically increase their contributions in the following period, but the more they are punished the more they will increase it (EI-NA0 and NEI-A0). With free-rider anonymity however, low contributors neither automatically increase their contributions nor does punishment induce them to do so ($F = 1.01$, $p = 0.3664$). This is in contrast to high contributors, where with free-rider anonymity the higher a subject's contribution is relative to others, the more they will revise it downwards in the following period. The combination of punishment being unable to induce low contributors to increase their contributions and high contributors subsequently contributing less, explains why the overall level of cooperation is lower in free-rider anonymity than the other treatments, supporting Hypotheses 1 and 3.

5 Conclusion

Our objective in this paper was to question how much the success of punishment in sustaining social norms relied on there being no legitimate but unobservable

[22] One could analyze the change in a subject's contribution from period t to the next period he is included in. However, the gap between the decisions could introduce confounding factors. Note that considering only subjects taking part in two consecutive periods does not reduce the number of observations dramatically.

[23] Punishment may still be driving this effect as if punishment is well targeted at lower contributors, the absolute deviation will be capturing this effect.

reason to contribute zero. In settings where it was possible to infer who was free-riding we obtained the familiar result that punishment leads to sustained cooperation (e.g. Fehr and Gächter 2000). When it was not possible to infer who was free-riding this was not observed. Instead, free-rider anonymity caused an adverse punishment profile where subjects who did not have the opportunity to contribute, and those who contribute more than others, received as much punishment as free-riders. This led to the breakdown of cooperation. One might have hoped that the inefficacy of punishment due to imperfect monitoring (Carpenter 2007b; Bornstein and Weisel 2010; Grechenig et al. 2010) could be overcome if subjects determine the degree of monitoring imperfection themselves, our results suggest otherwise.

In interpreting our results we would not want to argue that punishment cannot sustain cooperative social norms. It clearly can, and does so in three of our treatments. Rather we would argue that it is vital to better understand the consequences of free-rider anonymity, and other forms of imperfect monitoring, in sustaining social norms. Free-rider anonymity seems common in many real world settings and if this does create problems in sustaining social norms then we need to evaluate the implications of this. This raises many interesting issues. For example, can people find effective ways to signal legitimate reasons to not contribute, or to find out who lacks legitimate reasons to not contribute? What motivations drive the punishment patterns we observe under free-rider anonymity? Can some free-rider anonymity be important in curtailing over-zealous punishment (Hugh-Jones and Reinstein, 2009)?[24] Finally, what happens in related contexts where the punishment of one individual spills over onto, potentially innocent, other group members or in settings where it is easy to monitor extreme free-riding but more difficult to monitor moderate free-riding?

References

[1] Anderson, C. M. and 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.

[2] Bornstein, G. and Weisel, O. (2010) "Punishment, cooperation and cheater detection in "noisy" social exchange," *Games*, 1: 18-33.

[3] Broberg, T., Ellingson, T. and Johannesson, M. (2007) "Is generosity involuntary?" *Economic Letters*, 94: 32-37.

[4] Carpenter, J. P. (2007a) "The demand for punishment," *Journal of Economic Behavior and Organization*, 62: 522-542.

[24] Although not reported, earnings in our experiment were the lowest in the baseline treatment as subjects punished excessively. The intermediate treatments are perhaps normatively most desirable since they lead to stable cooperation and reasonable payoffs. However it is difficult to draw conclusions from earnings since the benefits of cooperation via punishment typically take some time to emerge (Gächter et al. 2008).

[5] Carpenter, J. P. (2007b) "Punishing free-riders: How group size affects mutual monitoring and the provision of public goods," *Games and Economic Behavior*, 60: 31-51.

[6] Carpenter, J. P., Matthews, P. H. and Ong'ong'a, O. (2004) "Why punish? Social responsibility and the enforcement of prosocial norms," *Journal of Evolutionary Economics*, 14: 407-429.

[7] Carpenter, J. P., Bowles, S., Gintis, H. and Hwang, S-H. (2009) "Strong reciprocity and team production: Theory and evidence," *Journal of Economic Behavior and Organization*, 71: 221-232.

[8] Casari, M. and Luini, L. (2010) "Peer punishment in teams: expressive or instrumental choice?" *Experimental Economics*, forthcoming.

[9] Charness, G. and Dufwenberg, M. (2006) "Promises and partnership," *Econometrica*, 74(6): 1579-1601.

[10] Cinyabuguma, M., Page, T. and Putterman, L. (2006) "Can second-order punishment deter perverse punishment?" *Experimental Economics*, 9: 265-279.

[11] Dana, J. Cain, D. M. and Dawes, R. M. (2006) "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and Human Decision Processes*, 100: 193-201.

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

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

[14] Falk, A., Fehr, E. and Fischbacher, U. (2005) "Driving forces behind informal sanctions," *Econometrica*, 73(6): 2017-2030.

[15] Falk, A., Fehr, E. and Fischbacher, U. (2008) "Testing theories of fairness – Intentions matter," *Games and Economic Behavior*, 62: 287-303.

[16] Fehr, E. and U. Fischbacher (2004) "Third-party punishment and social norms", *Evolution and Human Behavior*, 25: 63-87.

[17] Fehr, E. and Gächter, S. (2000) "Cooperation and punishment in public good experiments," *American Economic Review*, 90(4): 980-994.

[18] Fehr, E. and Gächter, S. (2002) "Altruistic Punishment in Humans", *Nature*, 415: 137-140

[19] Gächter, S., Renner, E. and Sefton, M. (2008) "The long-run benefits of punishment," *Science*, 322: 1510.

[20] Gächter, S. and Herrmann, B. (2009) "Reciprocity, culture and human cooperation: previous insights and new cross-cultural experiment," *Philosophical Transactions of The Royal Society B*, 364: 791-806.

[21] Gächter, S. and Herrmann, B. (2010) "The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia," *European Economic Review*, forthcoming.

[22] Grechenig, K., Nicklisch, A. and Thöni, C. (2010) "Punishment despite reasonable doubt – A public goods experiment with uncertainty over contributions," *Preprints of the Max Planck Institute on Collective Goods Bonn*, 2010/11.

[23] Gürerk, Ö., Irlenbusch, B. and Rockenbach, B. (2006) "The competitive advantage of sanctioning institutions," *Science*, 5770: 108-111.

[24] Herrmann, B. Thöni, C. and Gächter, S. (2008) "Antisocial punishment across societies," *Science*, 319: 1362-1367.

[25] Hopfensitz, A. and Reuben, E. (2009) "The importance of emotions for the effectiveness of social punishment," *Economic Journal*, 119: 1534-1559.

[26] Hugh-Jones, D. and Reinstein, D. (2009) "Secret Santa: Anonymity, signalling and conditional cooperation," working paper.

[27] Lazear, E., Malmendier, U. and Weber, R. A. (2010) "Sorting, prices and social preferences," mimeo.

[28] Ledyard, J. (1995) "Public goods: A survey of experimental research," in Kagel, J. and Roth, A. (Eds.) *The Handbook of Experimental Economics*, Princeton University Press, Princeton.

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

[30] Nikiforakis, N. (2010) "Feedback, punishment and cooperation in public good experiments," *Games and Economic Behavior*, 68: 689-702.

[31] Nikiforkis, N. and Normann, H.-T. (2008) "A comparative statics analysis of punishment in public good experiments," *Experimental Economics*, 11: 348-369.

[32] van Prooijen, J.-W., Gallucci, M. and Toeset, G. (2008) "Procedural fairness in punishment systems: Inconsistent punishment procedures have detrimental effects on cooperation," *British Journal of Social Psychology*, 47: 311-324.

[33] de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A. and Fehr, E. (2004) "The neural basis of altruistic punishment," *Science*, 305: 1254-1258.

[34] Rabin, M. (1993) "Incorporating fairness into game theory and economics," *American Economic Review*, 83: 1281-1302.

[35] Reuben, E. and Riedl, A. (2009) "Public goods provision and sanctioning in privileged groups," *Journal of Conflict Resolution*, 53(1): 72-93.

[36] Sell, J. and Wilson, R. K. (1991) "Levels of information and contributions to public goods," *Social Forces*, 70(1): 107-124.

[37] Sousa, S. (2010) "Cooperation and punishment under uncertain enforcement," *CEDEX Discussion Paper* no. 2010-06.

[38] Sutter, M., Haigner, S. and Kocher, M. (2010) "Choosing the carrot or the stick? – Endogenous institutional choice in social dilemma situations," *Review of Economic Studies*, forthcoming.

[39] Tarui, N., Mason, C. F., Polasky, S. and Ellis, G. (2008) "Cooperation in the commons with unobservable actions," *Journal of Environmental Economics and Management*, 55: 37-51.

Table 1: Treatments

Number of groups per treatment	Exclusion Information (EI)	No Exclusion Information (NEI)
Announcing Zeros (A0)	11 (Baseline)	6
Not Announcing Zeros (NA0)	6	12 (Free-rider anonymity)

Figure 1: Contributions

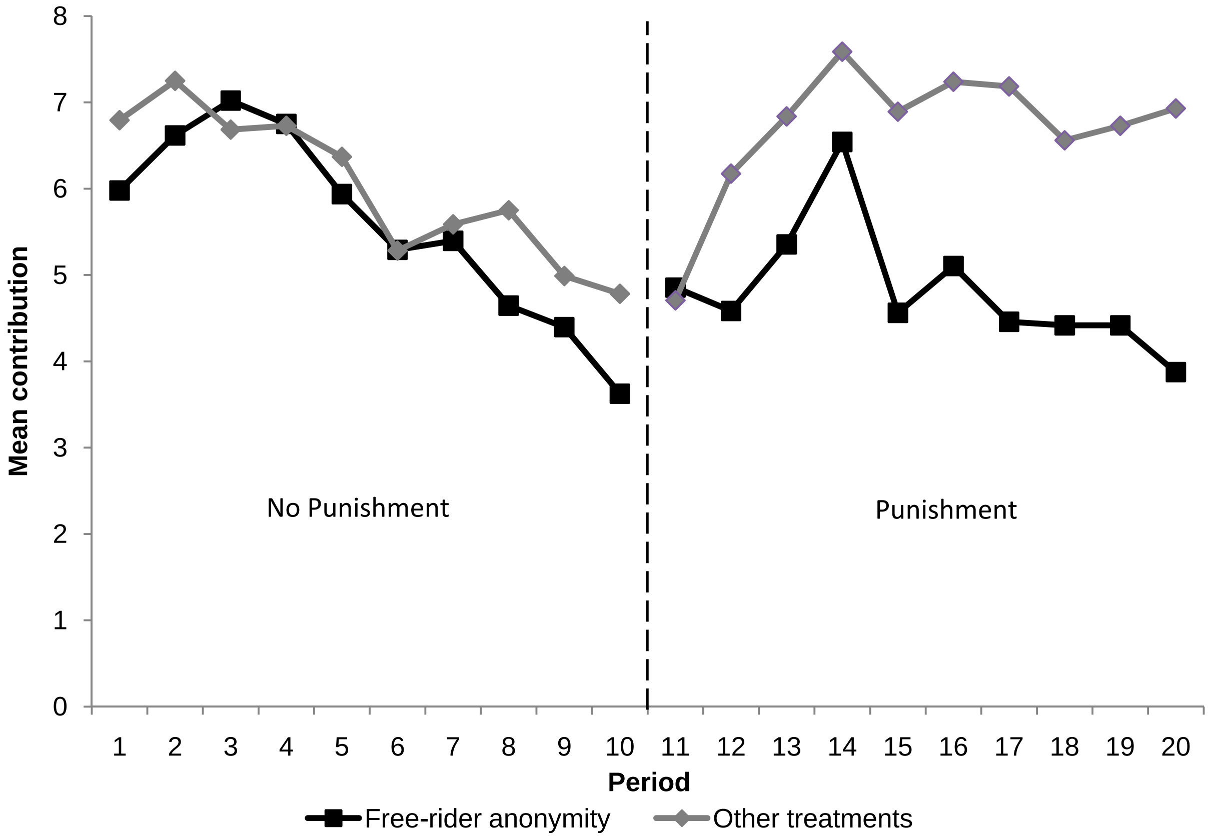


Table 2: Determinants of contributions

Censored Tobit: *Contribution*	Included only
*EI-NA0*No punishment*	0.4402 (1.586)
*NEI-A0*No punishment*	1.940 (1.324)
*NEI-NA0*No punishment*	-0.09138 (0.8985)
*EI-A0*Punishment*	-1.023 (1.274)
*EI-NA0*Punishment*	-0.1684 (1.586)
*NEI-A0*Punishment*	1.140 (1.988)
NEI-NA0*Punishment	**-2.717***** **(1.077)**
Constant	6.022*** (0.6930)
Period fixed-effects	Yes
F-stat	4.82***
Left-censored observations	439
Right-censored observations	124
N	2764

Notes: Robust standard errors clustered by group in parentheses; * denotes a 10% significance-level, ** idem 5% and *** idem 1%; *EI-A0, EI-NA0, NEI-A0 and NEI-NA0* are treatment dummies; *No punishment* and *Punishment* are dummies indicating periods 1-10 and 11-20 respectively.

Figure 2: Share of included subjects contributing zero

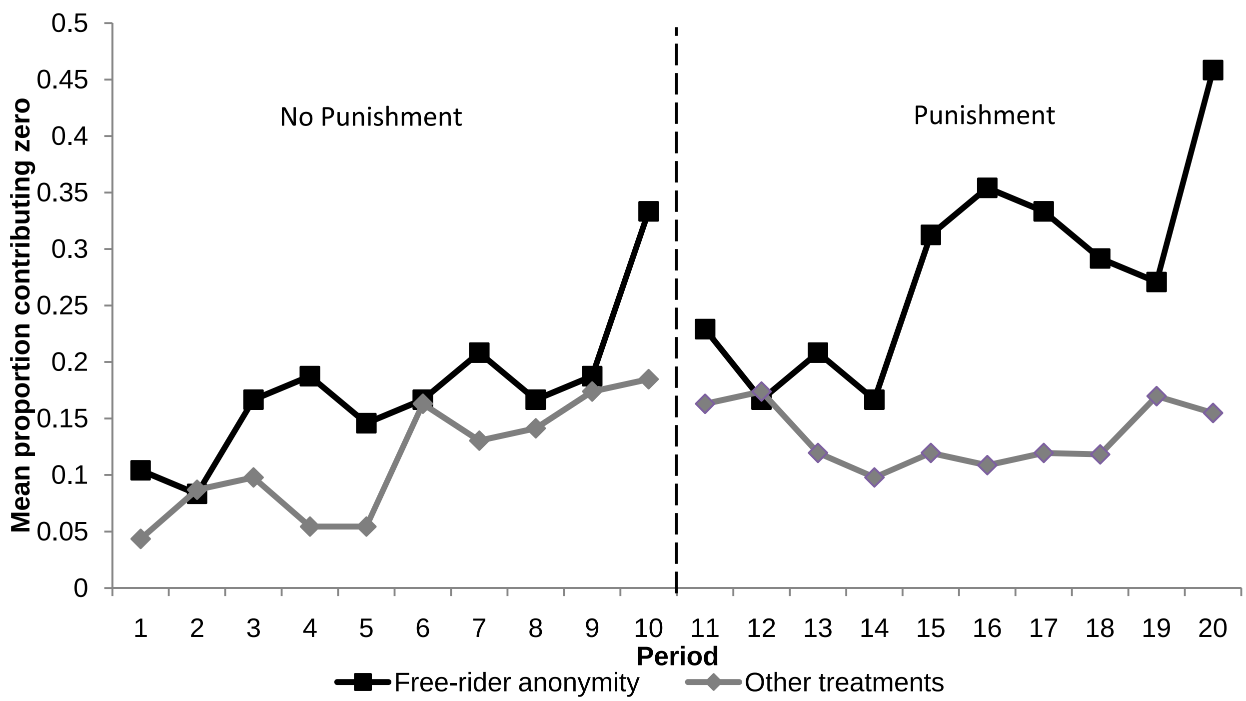


Figure 3: Punishment dispensed

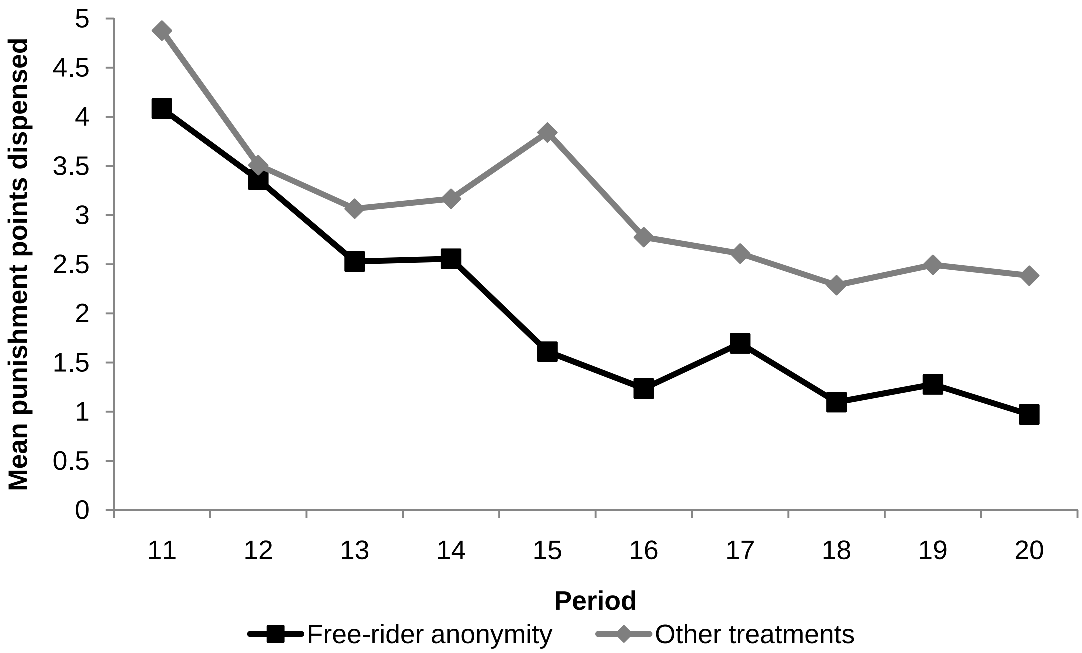


Table 3: Determinants of punishment dispensed

Censored Tobit: *Punishment dispensed*	All	Included only
EI-NA0	-2.355 (1.527)	-1.787 (1.500)
NEI-A0	-0.9178 (1.043)	-0.7293 (1.140)
NEI-NA0	**-3.403***** **(1.121)**	**-2.475**** **(1.014)**
Average contribution	0.1433 (0.1229)	-
Contribution variance	0.02237 (0.01545)	0.01132 (0.01543)
Excluded	0.4174 (0.6610)	-
Punisher's contribution	-	0.1237 (0.08014)
Contribution stage earnings	-0.1118 (0.07907)	0.01785 (0.09136)
Constant	6.143*** (1.986)	3.094 (2.480)
Period fixed-effects	Yes	Yes
F-stat	10.34***	4.94***
Left-censored observations	1009	652
Right-censored observations	1	1
N	2046	1364

Notes: *Excluded* is a dummy variable identifying excluded subjects; *Average contribution* was dropped in the included only model due to co-linearity with *Contribution stage earnings*.

Figure 4: Punishment by contribution deviation

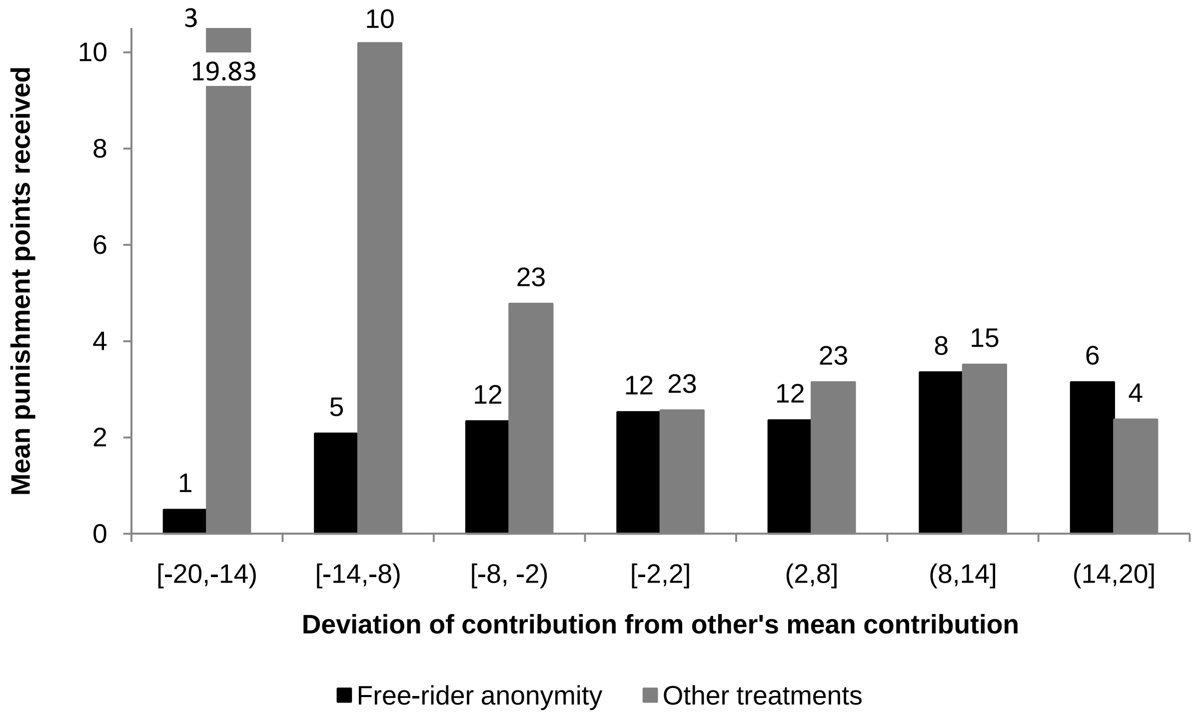


Table 4: Determinants of punishment received

Censored Tobit: *Punishment received*	Included only
Other's average contribution	-0.03422 (0.08357)
*EI-A0*Negative deviation*	1.089*** (0.1582)
*EI-NA0*Negative deviation*	0.7079** (0.2953)
*NEI-A0*Negative deviation*	0.9483*** (0.1204)
NEI-NA0*Negative deviation	**0.1617** **(0.09879)**
*EI-A0*Positive deviation*	0.2300* (0.1293)
*EI-NA0*Positive deviation*	0.1367 (0.1448)
*NEI-A0*Positive deviation*	0.3231*** (0.1100)
*NEI-NA0*Positive deviation*	0.1277 (0.1221)
*EI-A0*Free-rider*	-1.137 (1.331)
*EI-NA0*Free-rider*	-3.302** (1.650)
*NEI-A0*Free-rider*	-1.144 (0.8504)
NEI-NA0*Free-rider	**-1.338**** **(0.6604)**
Constant	3.714*** (0.7626)
Period fixed-effects	Yes
F-stat	15.48***
Left-censored observations	322
N	1364

Notes: *Negative deviation* is the absolute value of the negative deviation, if any, of a subject's contribution from the mean of others' contributions in the current period and is zero otherwise, idem *Positive deviation*; *Free-rider* is a dummy identifying included subjects that contribute zero.

Table 5: Determinants of the change in contributions

Censored Tobit: *Contribution change*	Low contributors	High contributors
*EI-A0*Absolute deviation*	0.3896*	-0.3199***
	(0.2236)	(0.1180)
*EI-NA0*Absolute deviation*	0.09471	-0.2829
	(0.2083)	(0.2251)
*NEI-A0*Absolute deviation*	0.07822	-0.1288
	(0.1970)	(0.1532)
NEI-NA0*Absolute deviation	**0.1911**	**-0.4675*****
	(0.1423)	**(0.1598)**
*EI-A0*Punishment received*	0.03388	-0.2833**
	(0.1608)	(0.1255)
*EI-NA0*Punishment received*	0.4221***	0.006008
	(0.1104)	(0.1819)
*NEI-A0*Punishment received*	0.4246**	-0.3619**
	(0.1969)	(0.1742)
NEI-NA0*Punishment received	**0.1577**	**0.07328**
	(0.1558)	**(0.1217)**
Constant	-0.7262	-0.4384
	(0.5647)	(0.6740)
Period fixed-effects	Yes	Yes
F-stat	2.84***	9.89***
Left-censored observations	118	44
Right-censored observations	12	27
N	430	368

Notes: *Absolute deviation* is the absolute difference between a subject's contribution in period *t* and the mean of others' contributions in period *t*; *Punishment received* measures the number of punishment points received by the subject in period *t*.

Appendix: Experiment Instructions

Notes: Instructions in parentheses [.] vary by treatment; screenshots are from the EI-NA0 treatment.

This is a computerised decision making experiment. Following the instructions carefully and making good decisions will earn you more money. You will be paid in cash at the end of the experiment.

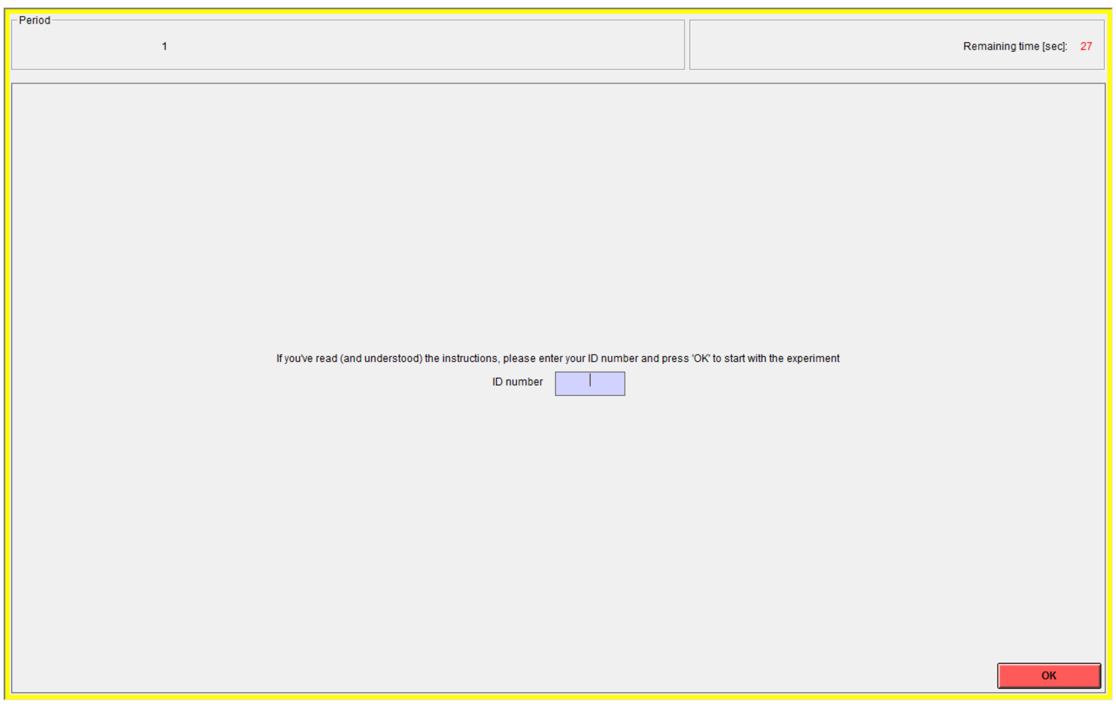
These instructions describe the decisions that you will be asked to make and how your earnings depend on these decisions. You may refer to these instructions at any time during today's session. **You must not communicate with other participants**. If anything is unclear at any stage, please put your hand up and we will come to you.

We will refer to “points” rather than Pounds throughout, where **1 Point = 2p or £0.02**.

At the start of the experiment, all subjects will be **randomly divided into groups of 6 participants**. The groups will stay the same throughout the experiment.

The experiment lasts for **20 periods**. The first 10 periods consist of 1 stage, the last 10 of 2 stages. At the end, you will be paid the amount you earned in total over the 20 periods.

Neither the experimenters nor the other participants will be able to connect you to your decisions. You will make your decisions alone at a computer and only be identified by your ID and participant numbers. By your computer you will find a 3-digit number, your ID number. When you finish reading these instructions, you must enter your ID number on the first screen and press “OK” to confirm you understand the instructions and are ready to begin.



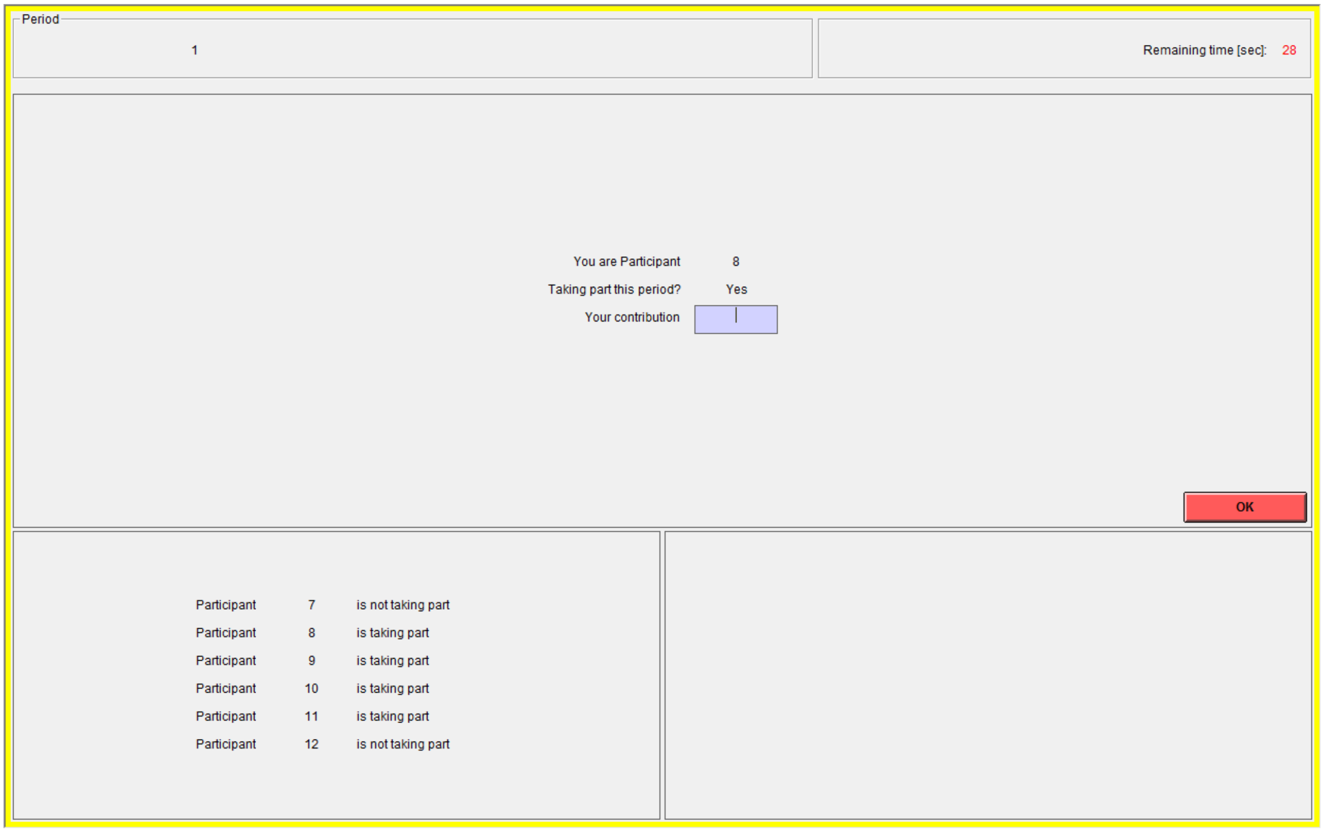
Each participant will be assigned a **participant number** which will not change throughout the session.

<u>Stage 1 (All periods)</u>

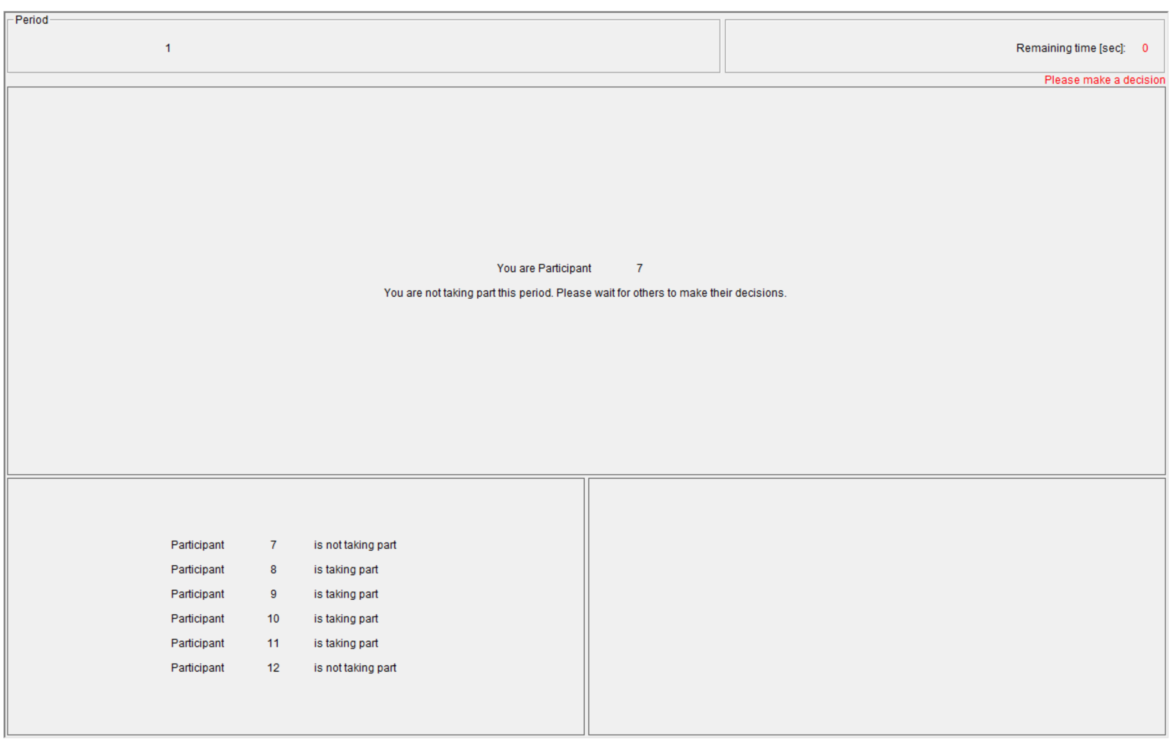
At the beginning of *each* period, each of the 6 members of the group receives **20 points**.

The computer will then **randomly select 4 of the 6 group members to take part in that period**. [You **will see which participants have been selected** in the lower left panel of the screen / You will only know whether or not you have been selected and **will not know which other participants have been selected]**

If you are selected to take part in a period, you will be asked how many of your 20 points you would like to contribute to the group and how many you would like to keep for yourself. You can **contribute any number of your points (from 0 to 20)** by entering it in the space provided and clicking “OK”. Contributions over 20 and decimal contributions will not be accepted. You will not know how many points others have contributed when you make this decision.



If you are not selected to take part in a period, you still receive the initial 20 points [and see who has been chosen to take part.] You must wait for the selected participants to make their contribution decisions.



Once all participants have made their contribution decisions earnings are calculated.

Your earnings from Stage 1 depend on **your contribution** and the **total number of points contributed** by the 4 people taking part in that period.

Stage 1 Earnings = 20 – Your contribution + 0.4*Total contributions by group

In other words, your Stage 1 earnings have two parts:

1. The points you keep for yourself (20 – Your contribution)
2. Earnings from group contributions (40% of Total group contributions)

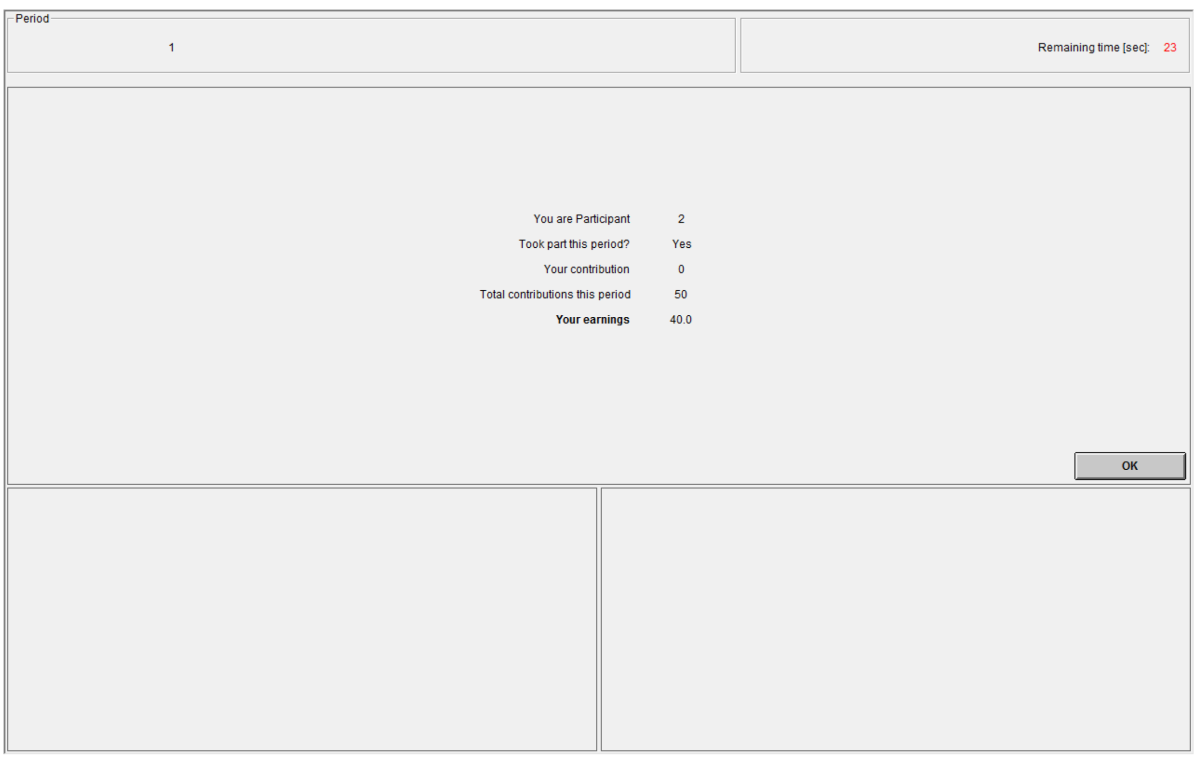
All group members' earnings are calculated in this way, including those who are not been chosen to take part in a particular period. This means that each group member receives the same earnings from group contributions.

Suppose total group contributions are 60 points, then each group member will receive 0.4*60 = 24 points as earnings from group contributions. If they are 9 points, then each group member receives 0.4*9 = 3.6 points as earnings from group contributions.

You may contribute as few or as many points as you like. **Each point you keep raises your earnings by 1 point. Each point you contribute increases every group member's earnings (including your own) by 0.4 points. Similarly, each point contributed by another group member raises your earnings by 0.4 points.**

In addition to the 20 points you are given at the start of each period, you will be given a one-off 25 points at the start of the experiment. This can be used to cover any losses you make in the experiment. **You can always evade losses with certainty through your own decisions**. This one-off payment will not be used to calculate period earnings; it will be added right at the end of the experiment.

In periods 1-10 this is the only stage. Once you have read the total contributions made by the group and your earnings, click "OK".



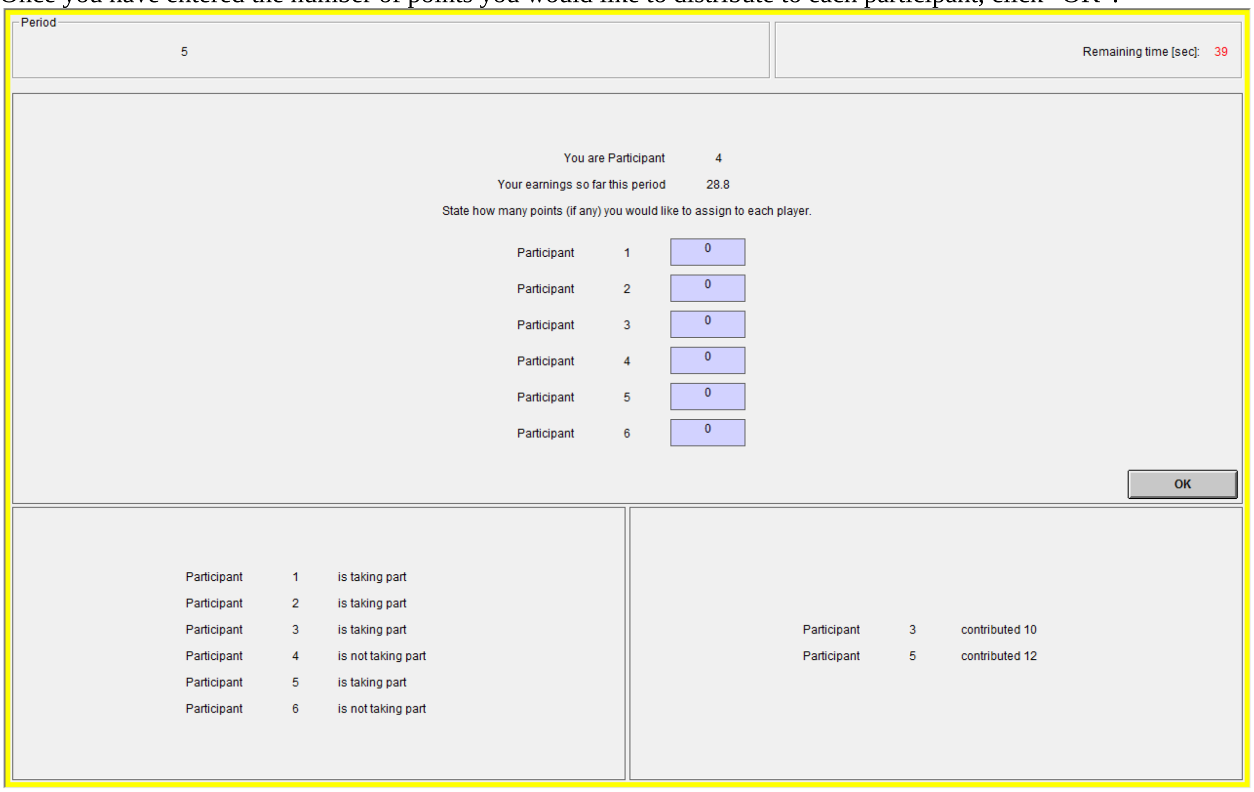
<u>Stage 2 (Periods 11-20 only)</u>
Periods 11-20 have a further stage before your period earnings are calculated. In the second stage, [you will be reminded who was chosen to take part (bottom-left panel of screen)], you will be [**informed of each group member's contribution** (bottom-right panel of screen) / **told the participant number and contribution** of those who made a **positive contribution; those who contributed zero will not be shown** (bottom-right panel of screen). Similarly, if you made a positive contribution in Stage 1, other participants will see this contribution; if you contributed zero, this will not be shown to them.] You can now choose to **reduce or leave equal** the earnings of **each** member.

To reduce another participant's earnings you have to **distribute points**. Each point will **cost you 1 point and reduce the earnings of the person you assign the point to by 3 points**. If you choose 0 for a particular participant, their earnings do not change.

You may distribute as many or as few points as you like, provided you can afford to pay the cost of distributing points from Stage 1 earnings. Decimal points are not accepted.

Note that your earnings may also be reduced as **all other group members are also given the opportunity to distribute points**, including those not chosen to take part in Stage 1.

Once you have entered the number of points you would like to distribute to each participant, click "OK".



Your earnings for the period (Stage 1 and Stage 2) are then:

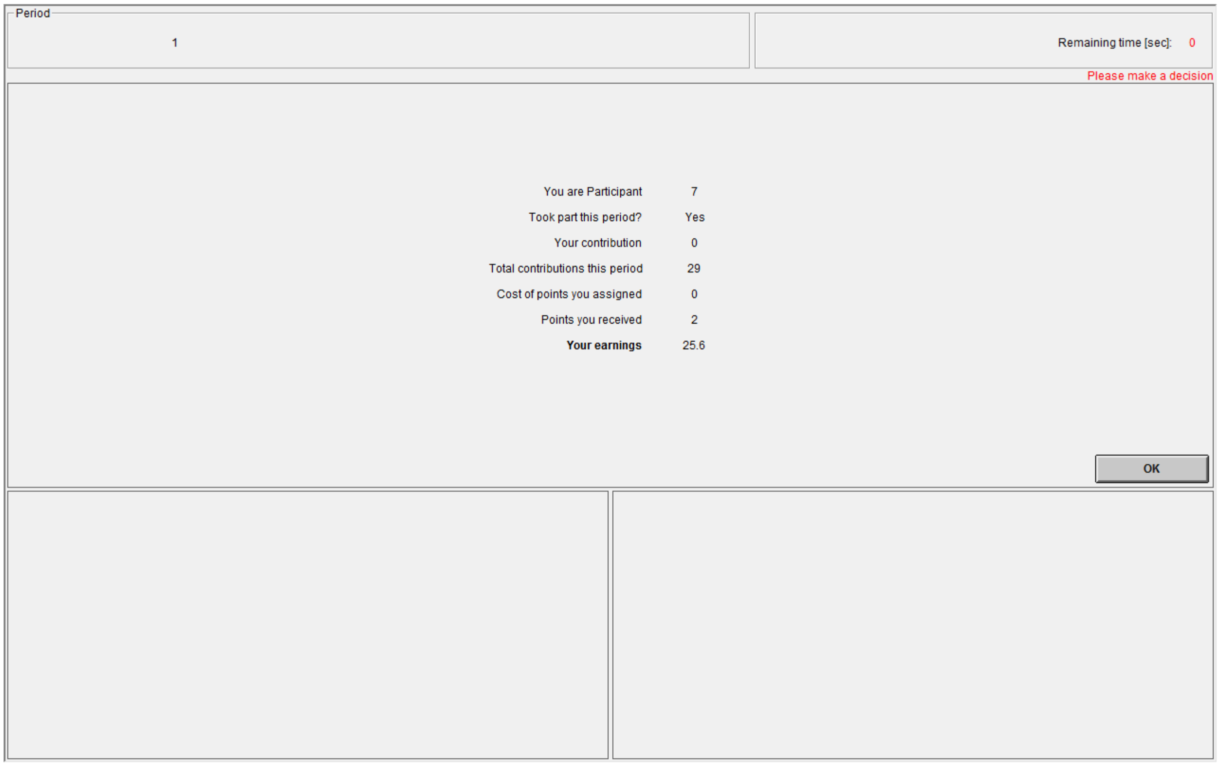
Earnings = Earnings from Stage 1 – Points you give – 3*Points received

In other words, in Stage 2 your earnings from Stage 1 are reduced by:
1. The cost of the reductions that you make to others' earnings
2. The amount that other participants choose to reduce your earnings by

So if you earned 30 points from Stage 1, distributed 4 points and received 2 points: Your Earnings = 30 – 4 – 3*2 = 20.

It is possible that your period earnings can become negative if you receive many points from others. You can use the one-off 25 points to pay for this.

Once all participants have decided how to assign points, you will see your earnings for that period and how they were calculated. The next period begins when all participants have clicked "OK".



To summarise, a period will proceed as follows:

1. You are given 20 points and the computer randomly selects whether you are 1 of the 4 people taking part in the period. [You see which participants are taking part / You aren't told which participants are taking part.]
2. If selected to take part, you may contribute any number of your 20 points (0 to 20).
3. (Periods 11-20 only) After contribution decisions are made, you are informed of all contributions [except people who contribute zero.] You can then use your earnings so far to reduce others' earnings. It costs 1 point to reduce a participant's points by 3.
4. Your period earnings are then calculated as follows:
   a. *Periods 1-10*
      Your Earnings = 20 points – Your contribution + 0.4*Total group contributions
   b. *Periods 11-20*
      Your Earnings = 20 points – Your contribution + 0.4*Total group contributions
      – Points you distribute – 3*Points you receive
5. Earnings are then displayed.

Note that you will receive the initial 20 points at the start of **each** period. At the end of the experiment you will be paid your total earnings across all 20 periods (1 point = 2p or £0.02) plus the 25 points one-off payment. Therefore if you earn 600 points over the 20 rounds, your total points would be 625 and you would be paid 625*0.02 = £12.50 in cash.

It is **important** you understand the procedures and how earnings are calculated. If you have any questions at any stage raise your hand and we will come to you. If you are ready to begin, enter your ID number. The experiment begins once all ID numbers are entered.