Accepted Manuscript

Varying experimental instructions to improve comprehension: punishment in public goods games

Abhijit Ramalingam , Antonio J. Morales , James M. Walker

PII: S2214-8043(18)30055-7
DOI: 10.1016/j.socec.2018.01.008
Reference: JBEE 327

To appear in: *Journal of Behavioral and Experimental Economics*

Received date: 22 June 2017
Revised date: 31 January 2018
Accepted date: 31 January 2018

Please cite this article as: Abhijit Ramalingam , Antonio J. Morales , James M. Walker , Varying experimental instructions to improve comprehension: punishment in public goods games, *Journal of Behavioral and Experimental Economics* (2018), doi: 10.1016/j.socec.2018.01.008

Highlights

- Compares effects of two sets of instructions on behaviour in public goods games
- Detailed instructions are associated with higher comprehension levels, measured by decision times
- Without punishment, instruction format does not affect public goods contributions
- With punishment, contributions are higher with longer, more explicit instructions
- Detailed instructions are associated with better targeting of low contributors for punishment

Varying experimental instructions to improve comprehension: punishment in public goods games

Abhijit Ramalingam [a*], Antonio J. Morales [b], James M. Walker [c]

[a] School of Economics and Centre for Behavioural and Experimental Social Science, University of East Anglia, Norwich NR4 7TJ, UK

[b] Facultad de Economía, Universidad de Málaga, Spain

[c] Department of Economics and the Ostrom Workshop, Indiana University, USA

Abstract

We provide evidence that more explicit instructions can affect behaviour in a public goods game with punishment. Instructions that highlight the positive externality associated with public goods contributions and provide more examples improve subjects' comprehension levels, as measured by shorter decision times in the experiment. They also lead to higher contribution levels in games with punishment opportunities, linked to better targeting of punishment.

Keywords: public goods, experiment, instructions, contributions, punishment, methodology

JEL codes: C72, C91, C92, H41

[*] Corresponding author: a.ramalingam@uea.ac.uk, abhi.ramalingam@gmail.com, Tel: +44-1603-597382

1. Introduction

There is ample evidence that the format in which instructions are presented to students can significantly influence the degree to which learning can be facilitated. Chandler and Sweller (1991) develop the cognitive load theory to argue that information should be presented in ways that do not impose heavy cognitive loads – otherwise hampering learning. Modern versions differentiate three kinds of cognitive loads: intrinsic (the complexity of the matter itself), germane (effective) and extraneous (irrelevant). While these principles have guided the design of instructions in (educational) psychology and educational research (see De Jong, 2010), there has been little research in experimental economics that focuses explicitly on how alternative instructions impact subjects' understanding of the decision task.

A large amount of attention has, instead, focused on the effects of decision frames on behaviour. Since the seminal work by Tversky and Kahnemann (1981), plenty of experimental work in economics has analysed how the description of decision problems and strategic situations affects people's perception of the situation, and their choices and behaviour. As relevant examples for our research, in a public goods game, Andreoni (1995) studies the differences in contributions in positive vs negative frames and, more recently, Cubitt et al. (2011) and Ramalingam et al. (2017) study differences in contribution and punishment behaviour in one-shot and repeated provision vs appropriation games.

In this paper, we study the effect of the instruction format not on subjects' perception of a situation but on subjects' comprehension of the situation, designing instructions to increase subject comprehension of the incentive structures of the laboratory decision setting. Ultimately, the goal is to study whether higher comprehension affects behaviour.

Recent work documents that instruction format can have significant effects on comprehension levels of subjects in economics experiments. In a repeated linear public goods experiment, Bigoni and Dragone (2012) (henceforth BD12) identify two factors that influence the effectiveness of experimental instructions – their length and subjects' active involvement. They find that short on-screen instructions worsen comprehension (measured by the number of wrong answers, and the time taken to answer a pre-experiment quiz), while instructions of a similar length that required subjects to actively solve examples during the instruction stage were found to increase comprehension levels.[1] However, BD12 do not find

[1] They did not run treatments with online long instructions or long instructions that required active input.

evidence that differences in comprehension levels are associated with differences in contributions. They observe the usual pattern of behaviour (see Chaudhuri, 2011), with declining contributions across treatments.

While these results regarding behaviour might be reassuring at first glance, there are reasons for further investigation into this aspect of the experimental methodology. A possible reason for the finding of BD12 is the relatively simple nature of the public goods game where a participant makes one decision - contribution - in each round. Instruction length and format may thus not have serious implications for behaviour in such a setting. However, combined with previous evidence that some subjects' decisions appear to be at least partially linked to confusion (see Andreoni, 1995), it is plausible that lack of comprehension would have more noticeable effects on behaviour in more complicated settings.[2]

As suggested by Cognitive Load Theory, intrinsic cognitive load is higher in more complex settings. However, instructions designed to improve understanding of the more complex setting are likely to reduce the germane load on experimental subjects. It is plausible that such instructions have an impact on behaviour because they reduce the overall cognitive load. Our study presents evidence that supports this conjecture.[3]

One of the most studied results in the social dilemmas literature is the ability of groups to use peer punishment to govern themselves, i.e., to raise cooperation levels and sustain these higher levels over time (for instance, Ostrom et al., 1992 and Fehr and Gächter, 2000). Successful punishment is associated with three factors: (i) sufficient punishment to render the threat credible, (ii) punishment targeted at free-riders, and (iii) punishment of low contributors not being crowded-out by anti-social punishment of high contributors (see Hermann et al., 2008 and Rand et al., 2010).

Arguably, therefore, the punishment game is more complex than a simple public goods game without the option to sanction one another. First, there are more decisions to make in the game with punishment. Second, the punishment decision involves more complicated reasoning on the part of subjects in identifying who and how much to punish. If, as has been shown, instruction length has significant effects on comprehension, we are more likely to see

[2] BD12 note that their conclusions may "depend on the complexity of the task" (p. 463).

[3] We are aware of one other work that looks at different instruction formats, based on Cognitive Load Theory. Kirmes (2014) provides preliminary evidence from a pilot experiment that instructions with stick-figure illustrations (that help understanding by reducing the germane load) reduce decision times in public goods and market entry games. However, given that these results are preliminary, the findings are not conclusive.

effects on outcomes in more complicated settings, i.e., in the punishment game. Given the increasing complexity of some experiments in recent times, and the resulting wide variety of instructions used, it is extremely important, and timely, to investigate the potential effects that instruction format and complexity may have on subject comprehension levels and behaviour.

In this study we examine the effect of instruction format on decisions in two environments that differ in complexity; experimental public goods games with and without punishment opportunities. Thus, our treatments vary along two dimensions – availability of punishment and format of instructions – to implement a 2 × 2 design. In all four treatments, subjects played a repeated linear public goods game using the Voluntary Contributions Mechanism (VCM). In two treatments, subjects played only the VCM. In the remaining two treatments, subjects played a two-stage game where the first stage was the VCM. In the second stage, subjects could use their earnings from the first stage to sanction each other (as in Gächter et al., 2008).

In one pair of VCM and punishment treatments, subjects were given short instructions while in the other pair, subjects were given longer instructions that made explicit the positive externality associated with contributions to the public good. Versions of both sets of instructions have been widely used. The longer instructions were based on Gächter et al. (2008) while the shorter instructions were based on Fatas and Mateu (2015). We used important elements from the instructions in these papers and further adapted them to reflect our experimental parameters.

While punishment experiments that use the longer instructions (including Gächter et al., 2008) have generated sustained increases in contributions, Fatas and Mateu (2015) find only modest increases in contributions. To the best of our knowledge, there has been no systematic investigation of such differences in punishment behaviour, or of reasons for the differences. It is thus not yet clear if, or how, the length and format of instructions may have an influence on observed differences across studies examining the punishment institution.

The two instruction formats in our study differed in important respects. First, the shorter instructions were one and two pages long respectively for the games with and without punishment while the corresponding longer instructions were three and four pages long. Second, the shorter instructions had only two solved examples each while the longer instructions had three examples each. Third, and perhaps most important, the longer

instructions made salient the positive externality inherent in the public goods game while the shorter instructions stopped with the description of the game and calculation of payoffs.

Our results lend support to the findings in BD12. As in BD12, we find that the shorter instructions do negatively affect comprehension levels in the VCM. When given shorter instructions, subjects took significantly longer to make contribution decisions in each round. As in BD12, we find that that average contributions start higher in the VCM sessions with longer instructions, but averages across all decision rounds are similar and show the common trend of declining contributions across all decisions rounds.

As in the VCM, comprehension levels were significantly lower in the shorter instructions punishment treatment. Subjects took significantly longer to make contribution and punishment decisions when they were given short instructions. What is different is that instruction length is associated with significant differences in behaviour in the punishment game. We find that when given the short instructions, groups were less successful in raising contribution levels. At best, they were able to stem the decline observed in the VCM across decision rounds. On the other hand, groups that received the longer more explicit instructions raised contributions to 75% of endowment and were able to sustain this higher level throughout the game. This is driven by differences in punishment behaviour – low contributors were targeted for punishment much more often.

The rest of the paper is organised as follows. Section 2 details our experimental design and presents the crucial differences between instruction formats. Section 3 presents our hypotheses, Section 4 presents our results, and Section 5 concludes. Appendix A contains the experimental instructions for all our treatments. Appendix B contains the pre-experiment quizzes that subjects had to answer. Appendix C presents additional analysis that explores subjects' response times further, and heterogeneity among groups in the punishment treatments.

2. Experimental design

In all treatments, groups of four subjects played a repeated VCM game. Each player was endowed with 20 tokens which could be invested in an *Individual Project (IP)* or in a *Group Project (GP)*. A token invested in the IP yielded a return of one token and a token invested in

the GP yielded a return of 0.5 tokens for each of the four group members, i.e., MPCR = 0.5. The per-period payoff of player *i* (*i* = 1, 2, 3, 4) is given by

$$\pi_i(c) = (20 - c_i) + 0.5 \sum_{i=1}^{n} c_i \,,$$

where c_i is *i*'s contribution to the public good, and c is the contribution profile in the group. The Nash equilibrium, based on self- regarding preferences and the same belief for all other players, is for all players to contribute 0 tokens to the GP. The social optimum is for everyone to contribute all 20 tokens to the GP. The equilibrium and the optimum remain unchanged in finite repetitions.

Our treatments varied two dimensions, resulting in a 2 × 2 design: (1) whether or not there were punishment opportunities in the VCM game and (2) two types of instructions. In two conditions (VCM), subjects played only the VCM game described above. In two conditions (Pun), subjects played a two-stage game. The first stage was the VCM game. In the second stage, subjects could use their first-stage earnings to punish each other. As in Gächter et al. (2008), a player could use at most 5 tokens to sanction any other player. Thus, in total a subject could assign up to 15 punishment tokens in the second stage, or his/her stage 1 earnings, whichever was lower. The punishment technology was 1:3, i.e., one punishment token assigned cost the sender 1 token and the recipient 3 tokens. Per-period payoffs are now given by

$$\pi_i(c, p) = (20 - c_i) + 0.5 \sum_{i=1}^{n} c_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ij} - 3 \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ji} \,,$$

where p is the punishment profile in the group, p_{ij} is the number of punishment tokens assigned by player *i* to player *j* ($j \neq i$), and p_{ji} is the number of punishment tokens assigned by player j to player i ($j \neq i$). The Nash equilibrium in the punishment game is zero contributions and zero punishment by all players. The social optimum is full contributions and zero punishment by all players. Both remain unchanged in finite repetitions.

In both VCM and Pun treatments, we explore differences in subject comprehension and behaviour, contrasting behaviour when subjects read shorter/less explicit instructions versus

longer/more explicit instructions (online Appendix A).[4] Our longer instructions were designed to provide more information and details, with a view to improving subjects' understanding of the decision situation. The two sets of instructions differ in several ways. For brevity, we use the terms "LOW" (L) and "HIGH" (H) to refer to the targeted comprehension levels in, respectively, the shorter and the more detailed sets of instructions.[5] The differences between our short and long instructions are as follows.

In the HIGH instructions, to direct subjects' attention to the positive externality associated with public good contributions, the spill over was made more explicit. The LOW instructions simply described the calculation of earnings from the contribution stage, i.e., described the first payoff function above in words. They did not mention the implications of one's contributions for other group members. The HIGH instructions stated "For each token you allocate to the Group Project, you will earn 0.5 tokens. Each of the other three people in your group will also earn 0.5 tokens. Thus, the allocation of 1 token to the Group Project yields a total of 2 tokens for all of you together." They also stated "you will earn from your own allocation as well as from the allocations of others." The HIGH instructions used a bold font to emphasize key attributes of the decision settings.

The LOW and HIGH instructions contained, respectively, two and three solved examples each. The two examples in VCM-L illustrated the payoff consequence of different levels of contributions by the four subjects in the group. Pun-L used the same contribution examples as VCM-L, with additional examples tied to how a subject's earnings were affected by punishing others and by being punished. The examples were presented at the end of the instructions. Thus the examples were designed to combine illustrations of contributions to the public good and illustrations of subjects' opportunity to reduce the earnings of other group members (examples of the payoff implications of sending and receiving punishment). In the VCM-H and Pun-H instructions, one example illustrated the zero contribution outcome, one example the social optimum outcome, and the third an intermediate level of contribution by one subject when the other subjects contributed nothing. Thus, in the HIGH instructions, the examples also focused on highlighting the positive externality associated with contributions. In contrast to the LOW instructions, the examples were presented after the description of Stage 1 of the game (the contribution stage) in both HIGH treatments. In Pun-H, the

[4] The more detailed instructions and the data from VCM-H and Pun-H were used in Ramalingam et al. (2016).
[5] In BD12, their long instructions were designed to add repetition of elements of the shorter instructions. In our experiment, the differences are more nuanced. This is the reason we repeat the VCM treatment with both sets of instructions in our experiment.

instructions went on to describe the punishment stage. However, there were no further examples; unlike in Pun-L, there were no examples of punishment use and their implications for earnings. This reflects our assumption that a better understanding of the social dilemma, a result of highlighting the positive externality, is sufficient to improve subjects' comprehension of the use of punishment as well.

The objective was to investigate if instructions *can* lead to differences in behaviour. Hence, as noted above, we allowed our LOW and HIGH instructions to vary in more than one dimension. However, our conjecture is that the greater understanding of the game afforded by greater focus on the positive externality is most likely to improve comprehension levels and, thus, affect behaviour in the more complex punishment game.[6] Once an effect on behaviour has been established, future work will need to explicitly test this conjecture.

In all sessions, subjects were given printed instructions that were also read aloud by the same experimenter in every session. Subjects could ask clarifying questions at any point, which were answered privately. All subjects then had to *individually* answer a quiz that tested their understanding of the calculation of the payoffs in the game.[7] There were three questions in the quiz in the VCM treatments, each of which asked subjects to calculate their own payoffs and the payoffs of others in the group for a hypothetical contribution profile. One question presented the Nash equilibrium allocation, one the social optimum and the third one presented a situation in which the subject was the only non-contributor in the group while the others in the group contributed 100% of their endowments. Subjects had to answer the same three questions in the quiz in the punishment treatments. However, there were two additional questions, the first of which asked the cost of assigning punishment and the second the cost of receiving punishment. Importantly, the quiz was the same regardless of the length of the instructions. In particular, subjects answered the same quiz in VCM-L and in VCM-H and the same quiz in Pun-L and in Pun-H.

Subjects were given calculators that they could use during the quiz and the experiment. Subjects had to answer all questions correctly before the experiment could begin. At the end of the quiz, all subjects were given printed answers (with explanations) to all the questions on the quiz. They could ask further clarifying questions at this stage.

[6] The HIGH instructions, though longer, are likely to impose a lower cognitive load, as they have reduced the load associated with trying to understand the game. In the LOW instructions, subjects have to figure out the important aspects of the game for themselves, thus increasing their cognitive load.

[7] Thus, subjects were actively involved, and forced inputs were required of subjects in all treatments. See BD12.

In all treatments, subjects interacted in the same group for 20 rounds (partner-matching). After the contribution stage in each round, subjects were shown the individual contributions of all players in their group in descending order. They were also shown the total contribution to the GP in their group and their earnings from the IP and the GP. No subject identifiers were used in order to minimise the possibility of reputation formation.

In the punishment treatments, subjects could assign punishment tokens to other individuals in their group in the second stage. Subjects were then shown the *total* punishment they received and their earnings from the two stages of the round. They were not told the identity of the group member who punished them.

Table 1 summarises characteristics of each treatment and the number of independent observations (groups) and subjects in each.

Table 1. Summary of treatments

Treatment	Punishment?	Length of instructions	Inst. word count	No. of examples	No. of groups	No. of subjects
VCM-L	No	Short	503	2	11	44
Pun-L	Yes	Short	834	2	15	60
VCM-H	No	Long	1039	3	10	40
Pun-H	Yes	Long	1345	3	12	48
Total					48	192

The experiment was programmed in z-Tree (Fischbacher, 2007). All sessions were conducted at EssexLab by the same experimenter. Subjects were recruited from the undergraduate student body at the University of Essex and had not participated in a public goods experiment before this. To minimise timing effects, we ran different sessions for each treatment at different times of the day. No subject participated in more than one treatment. Sessions with short instructions lasted 70 minutes on average, while sessions with long instructions lasted 55 minutes on average.[8]

Earnings could not be carried forward for use in future rounds and subjects received a fresh endowment of 20 tokens in each round. Subjects were paid their earnings from all 20 rounds of the game. Token earnings were converted to Pounds at the rate of 60 tokens to £1 and subjects received an average payment of approximately £12 including a £2.50 show-up fee.

[8] This already indicates that comprehension levels, as measured by decision times, are higher in the HIGH treatments.

3. Hypotheses

Our conjecture is that the HIGH instructions will be associated with higher subject comprehension levels regarding the strategic incentives of the game and consequences for individual and group earnings. Better comprehension thus ought to allow subjects to work out the ramifications of their actions quicker. One implication of this would be that subjects would need less time to answer the pre-experiment quiz that tests their ability to calculate payoffs under hypothetical contribution scenarios. However, as mentioned earlier, two of the examples in the HIGH instructions also appeared on the quiz, while this was not the case in the treatments with the LOW instructions. Thus, looking at quiz times alone might present an incomplete picture of comprehension levels.

However, once the experiment began, subjects had to respond to actual decisions made by others in their groups. Better comprehension would imply shorter decision times here as well. Our proxy measure for understanding is thus the time taken by subjects to make contribution or punishment decisions.

Hypothesis 1: *The average time taken by subjects to make contribution and punishment decisions will be lower in treatments with the HIGH instructions*.

If subject comprehension is different across instruction formats, we should observe differences in contribution levels in the very first period, immediately after reading the instructions. BD12 find that short on-screen instructions are associated with lower comprehension levels. Further, they lead to lower average contributions to the public good *in the first round* than do the other three instruction formats they consider (their Figure 1, p.461).[9]

Hypothesis 2: *First-period contributions will be higher in VCM-H than in VCM-L, and in Pun-H than in Pun-L*.

Absent opportunities for punishment, BD12 do not find evidence of sustained differences in contributions across decision rounds associated with instruction format. Based on their

[9] BD12 do not test for differences, or analyse contributions in detail.

results, we hypothesize that instruction format will not lead to differences in contributions in our treatments that do not include punishment opportunities.

Hypothesis 3: *Aggregating across decision rounds, there will be no difference in average contributions between VCM-L and VCM-H.*

It has been shown in previous work that punishment opportunities raise contributions (e.g. Fehr and Gächter, 2000). This is due to the fact that low contributors are predominantly targeted for punishment. However, targeting of low contributors crucially hinges on subjects recognising that contributions are beneficial. That such targeting is not always observed (see Hermann et al., 2008 and Rand et al., 2010 for evidence on 'anti-social' punishment) suggests that the punishment game is more cognitively demanding. Further, related work finds that subjects are less successful at establishing a cooperative norm that can then be enforced by punishment in more complex games that involve heterogeneity (e.g., Reuben and Riedl, 2009 & 2013; Robbett, 2016) or interior equilibria (e.g., Cason and Gangadharan, 2015). If, as we hypothesise, the HIGH instructions are associated with higher comprehension levels of the positive externality associated with public good contributions, we would expect that subjects are more easily able to establish a cooperative norm in Pun-H than in Pun-L. Thus, we hypothesise that punishment is used more effectively to raise contributions in Pun-H than in Pun-L.

Hypothesis 4: *Average contributions will be higher in Pun-H than in Pun-L.*

4. Results

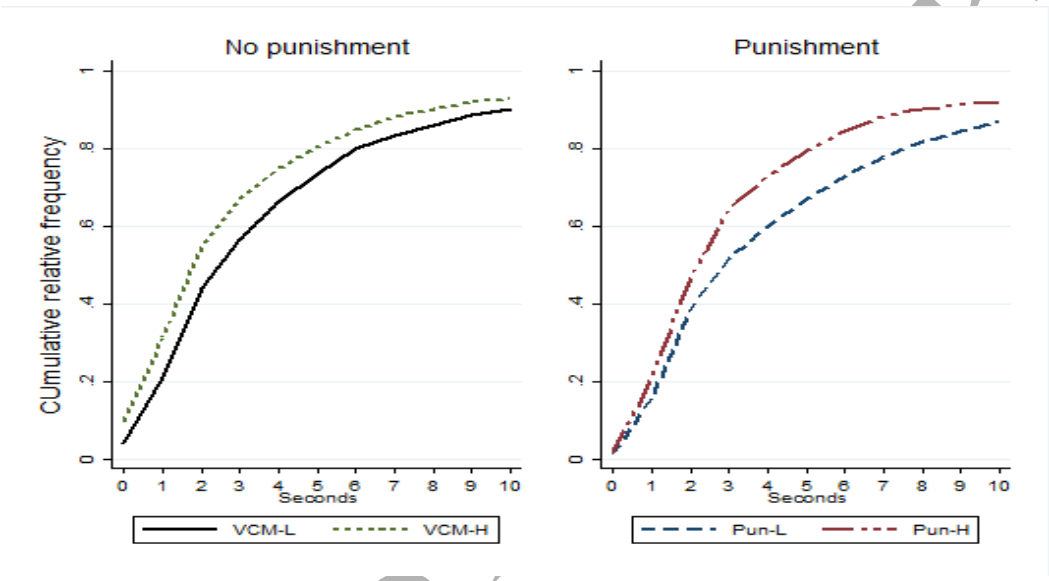
The presentation of results is organized around an exploration of the effect of differences in instruction length and content on comprehension levels in the different treatments, overall group performance, and punishment behaviour as an explanation for observed differences in behaviour. In all analyses, unless otherwise specified, Wilcoxon rank-sum tests are used to make pairwise comparisons between treatments and reported p-values are for two-sided tests.

4.1 Game comprehension – decision times

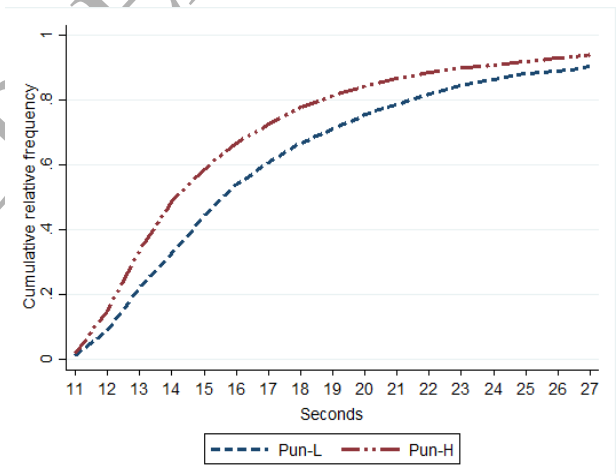
Figures 1(a) and (b) present the empirical CDFs of individual contribution and punishment decision times across all decision rounds.[10] To facilitate straightforward visual comparison, we display those observations with contribution (punishment) decision time up to 10 (27) seconds. At least 90% of all decisions were made within these time limits.[11]

Figure 1. Empirical CDF of individual decision times

(a) Contribution decisions



(b) Punishment decisions



[10] We also analyse the time taken by subjects to answer the pre-experiment quiz (Appendix C1). We find that subjects in the HIGH treatments take significantly less time answer the quiz than do those in the LOW treatments. One possible reason for why subjects perform better on the quiz in the treatments with the HIGH instructions could be that two of the examples in the instructions also appeared on the quiz.

[11] Figures C2 and C3 in online Appendix C1 present the corresponding figures using the entire range of observed decision times.

The figures show that the distributions of both contribution and punishment decision times for subjects who received the HIGH instructions displays first-order stochastic dominance over the distribution for subjects who received the LOW instructions, both with and without punishment opportunities.

Figure C4 in Appendix C3 presents empirical CDFs of contribution decision times comparable to Figure 2(a) for different contribution ranges. The figures show that the first-order stochastic dominance noted above is observed at zero contributions, low contributions (0 to 5 tokens), high contributions (16 to 20 tokens), and full contributions of 20 tokens. Thus, in treatments with the HIGH instructions, subjects took less time to make either low or high contributions. This suggests that the HIGH instructions were successful in increasing subjects' understanding of both the free-riding incentives and the benefits of high contribution levels, i.e., of the conflicting incentives in the game.[12]

In further analysis, Table 2 provides the average time taken to make contribution decisions (VCM-L and VCM-H) and contribution-punishment decisions (Pun-L and Pun-H) in a round. The first two columns present average decision times in the first round, and the last two columns present average decision times averaged over all 20 rounds. In relation to contribution or punishment decisions, because subjects made decisions within their groups, we treat a group as an independent observation. For each group, we first calculate the average of the variable for all 4 players in a round and then over all 20 rounds, thus resulting in one observation per group. The summary statistics in the table use this average for each group. The number of observations in each test is the number of groups in each treatment.

[12] We do not perform a similar analysis of punishment decision times. This is because punishment decisions are reactions to actual observed contributions in the group. A figure that breaks punishment levels into different ranges would have to control for the different contribution levels in the group in each round prior to each punishment decision. An empirical CDF cannot control for this.

Table 2. Mean decision time in seconds

		First round		All 20 rounds	
Treatment	**Obs.**	**Contribution**	**Punishment**	**Contribution**	**Punishment**
VCM-L	11	16.64	-	5.03	-
		(8.04)		(1.37)	
Pun-L	15	17.12	36.22	5.58	18.58
		(7.44)	(11.41)	(1.31)	(1.81)
VCM-H	10	12.70	-	3.72	-
		(5.59)		(1.11)	
Pun-H	12	20.60	30.52	4.48	16.67
		(13.04)	(4.82)	(1.39)	(1.45)

Figures in parentheses are standard deviations across groups within a treatment for the group experiment tasks. We dropped 6 instances (out of 3840) with a recorded contribution decision time of 99999, i.e., where subjects made decisions instantly.

Focusing on contribution behaviour in the first round and controlling for treatment conditions, Table 2 shows that subjects took less time to make contribution decisions in VCM-H than VCM-L, but more time in Pun-H than Pun-L. They also took less time to make punishment decisions in Pun-H than in Pun-L. None of the differences across treatments in the first round is statistically significant ($p > 0.10$ for all tests).[13]

Averaged over all rounds, subjects took less time to make contribution decisions in the treatments with the HIGH instructions. The differences are statistically significant for contribution decisions in both the VCM (n = number of groups in each treatment, $p = 0.0221$) and the punishment treatments (n = number of groups in each treatment, $p = 0.0570$). Subjects also took significantly less time to make punishment decisions in Pun-H than in Pun-L (n = number of groups in each treatment, $p = 0.0097$). Thus, while first round decision times are not very different across treatments, decision times drop over time in the treatments with the HIGH instructions.[14] We thus find support for Hypothesis 1.

Result 1: *The average time taken to make contribution and punishment decisions is significantly lower in the treatments with the HIGH instructions.*

[13] The standard deviation of punishment decision times in the first round is significantly lower in Pun-H than in Pun-L ($p = 0.0068$).
[14] Appendix C4 presents additional analysis of variation in decision times, within and across treatments.

4.2 Group public good contributions

Figure 2 presents mean group contributions across rounds in all treatments. Table 3 presents summary statistics of mean group contributions. As with decision times, an independent observation is a group of four subjects. For each group, we calculate the contribution to the public good in the first round, or the average contribution over all 20 rounds of the game.

Figure 2. Mean group contributions (group endowment = 80)

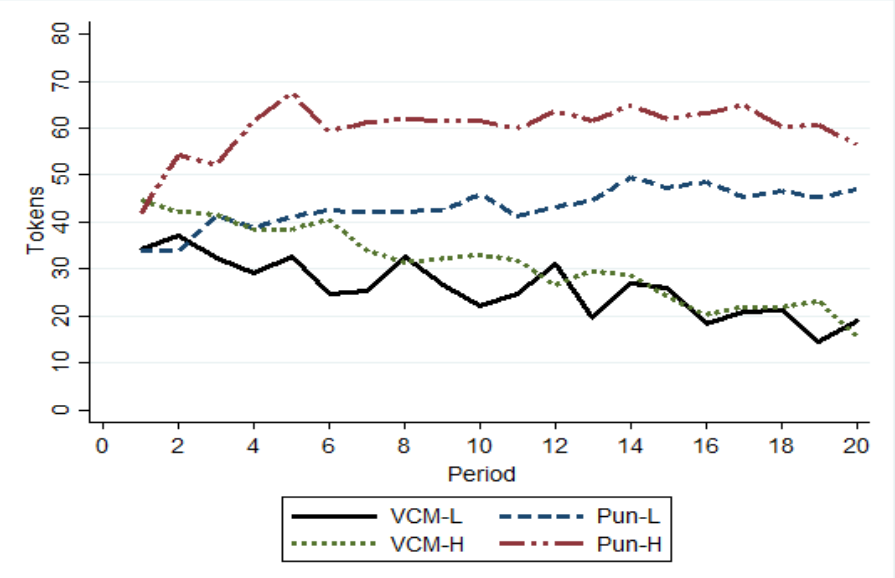


Table 3. Mean (st dev) group contributions (group endowment=80)

		Round	
Treatment	**Obs.**	**First**	**All 20**
VCM-L	11	34.27	25.98
		(10.53)	(11.44)
Pun-L	15	33.87	43.12
		(10.93)	(18.95)
VCM-H	10	44.7	31.01
		(14.28)	(14.21)
Pun-H	12	42.00	60.02
		(14.69)	(20.47)

Each group forms an independent observation. Each observation is the average of the relevant variable for that group, averaged over all 20 periods. Figures in parentheses are standard deviations.

Consistent with BD12, the different instruction formats lead to different contribution behaviour in the very first round; contributions start higher in the treatments with HIGH instructions, both with and without punishment. The difference is statistically significant in the absence of punishment opportunities, i.e., VCM-H vs. VCM-L (two-sided p = 0.0757, one-sided p = 0.0393) and marginally so in the presence of punishment opportunities, i.e., Pun-H vs. Pun-L (two-sided p = 0.1429, one-sided p = 0.0746). We thus find mild support for Hypothesis 2.

Result 2: *First-period contributions are higher in treatments with longer instructions, both with and without punishment. The difference is statistically significant in the VCM treatments and mildly so in the Pun treatments.*

Figure 3 shows that mean contributions start at just over 30 tokens in the LOW treatments and at about 45 tokens in the HIGH treatments. In the two VCM treatments, contributions follow the usual pattern and steadily decline to about 20 tokens, or 25% of group endowment, by round 20. Moreover, except for the initial few rounds, the dynamics over time are similar in both treatments. Table 3 confirms the general patterns observed in Figure 3. Parallel to BD12, average group contributions (over all rounds) in VCM-L are not significantly different from those in VCM-H (p = 0.4595). We find support for Hypothesis 3.

Result 3: *Average contributions in VCM-L and VCM-H are not significantly different.*

The decline in contributions is not observed in the two punishment treatments. Figure 3 and Table 3 show that overall average contributions in both punishment treatments are higher than in the two VCM treatments. Tests confirm that contributions in VCM-L are significantly lower than in Pun-L (p = 0.0274) and that contributions in VCM-H are significantly lower than in Pun-H (p = 0.0056).[15]

The figure shows a large difference in contribution levels between the two punishment treatments. In Pun-L, contributions stagnate at near initial levels throughout the 20 rounds. This is consistent with the result reported in Fatas and Mateu (2015) who also find that punishment leads to only a modest increase in contributions over initial levels. Contributions in Pun-H, however, rise to about 75% of endowment by round 5 and stay steady at that level for the remainder of the game. This pattern is similar to that widely documented in the

[15] Average contributions in VCM-L are also significantly lower than in Pun-H (p = 0.0011). However, average contributions in VCM-H are not significantly different than in Pun-L (p = 0.1077).

literature. Importantly, average (over all rounds) contributions in Pun-L are significantly lower than in Pun-H ($p = 0.0218$). We find support for Hypothesis 3.

Result 4: *Averaging over all 20 rounds, the HIGH instructions are associated with greater contributions in the decision environment that allows for punishment.*

Result 3 and Result 4 together imply that the boost to contributions from the introduction of punishment opportunities is greater in treatments with the HIGH instructions. To test for this, we calculate, within instruction formats, the average increase in group contributions over the corresponding average VCM group contribution (averaged over all groups). The average increase is 17.14 tokens in Pun-L and 29.01 in Pun-H. The difference is (mildly) significant ($p = 0.0877$).

4.3 Punishment use

The above discussion shows that the impact of instruction-format is mainly observed in the punishment treatments. A potential reason for this result may be related to differences in how punishment is used. We first investigate to what extent groups make use of punishment. Figure 3 presents average punishment used at the group level over time in the two treatments.

Figure 3. Mean group punishment

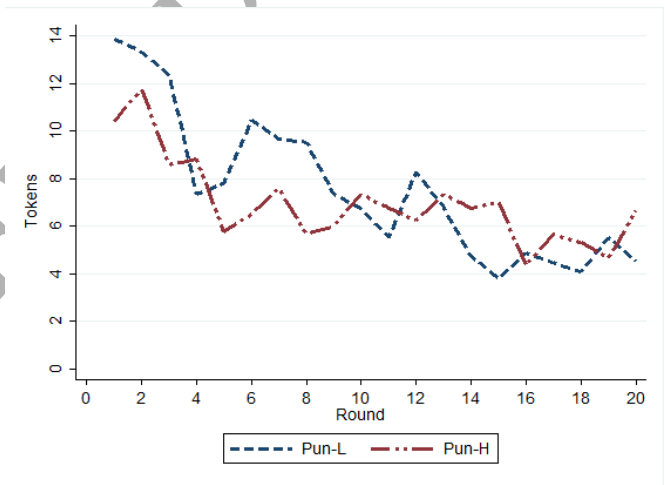


Figure 3 shows that mean group punishment is positive in all 20 rounds. Further, groups in Pun-L use more punishment, on average, than do groups in Pun-H in the earlier rounds of the game. However, by about round 10, there is no discernible difference in the level of punishment between the two treatments. Averaging over all 20 rounds, there is no significant difference in the average amount of punishment used in a round between the two treatments

(7.55 tokens in Pun-L vs. 6.96 tokens in Pun-H; n = number of groups in each treatment, p = 0.7883).[16, 17]

Although similar amounts of punishment were used in both punishment treatments, we do observe significant differences in outcomes between the two punishment treatments.[18] Evidence from prior studies suggests that *effective* punishment is associated with targeting low contributors and not targeting high contributors (blind revenge or anti-social behaviour). See, for example, the discussion in Hermann et al., 2008, Rand et al., 2010, and Ramalingam et al., 2016.

Because the HIGH instructions were designed to make the positive externality of contributions to the public good more explicit, and therefore the detrimental effect of free riding, we expect differences in the targeting of punishment across treatments.

Panel random-effects regressions were used to examine the amount of punishment received by an individual in a round. Since we are interested in seeing if low (high) contributors are targeted for punishment, we once again estimate separate regressions for negative and non-negative deviations. The model results are presented in Table 4 (first two columns). The amount of punishment received increases with the size of the negative deviation, but not with the size of the positive deviation of an individual's contribution from the average of the others in the group. Consistent with the aggregate finding reported above, instruction format does not significantly impact the *amount* of punishment received.

[16] The standard deviations were 5.96 in Pun-L and 6.37 in Pun-H.

[17] On average, 1.6 group members used punishment to reduce the earnings of other group members in a round in Pun-L, while the figure was 1.5 in Pun-H. The difference is not significant (p = 0.8835).

[18] Appendix C5 presents analysis of reactions to punishment – we do not find differences across instruction formats. Appendix C7 presents analysis that identifies 'successful' groups in Pun-L and Pun-H in terms of high contributions and explores their use of punishment.

Table 4. Determinants of received punishment

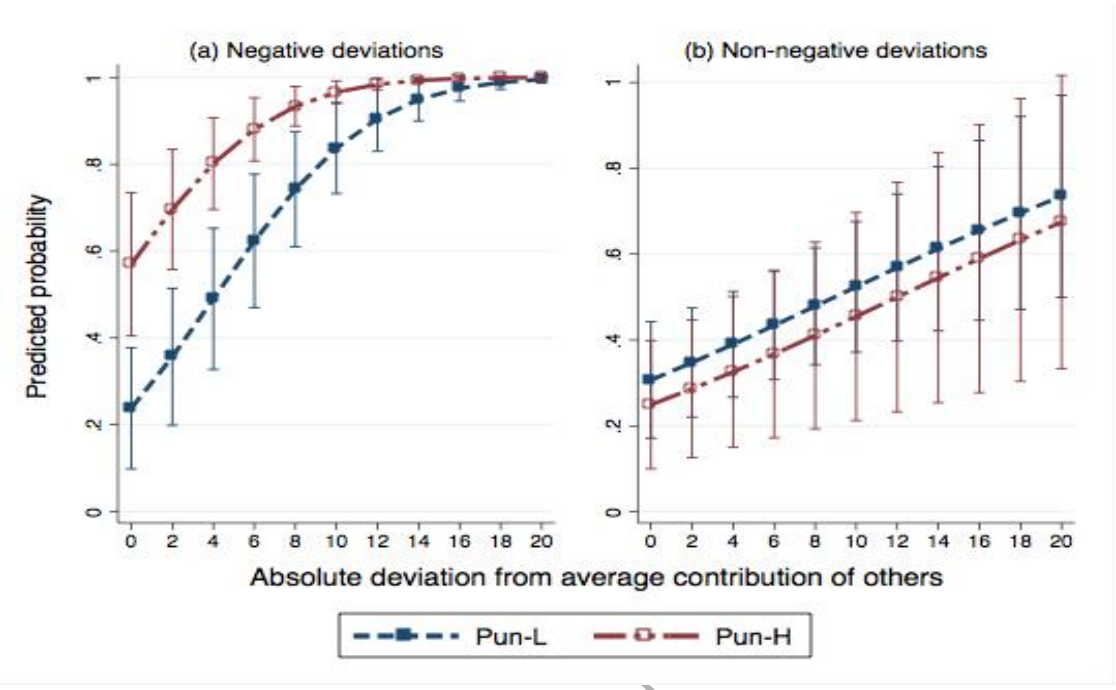
	Individual Random effects		Probit	
	Negative deviations	Non-negative deviations	Negative deviations	Non-negative deviations
Absolute deviation from average contribution of others	0.247***	0.045	0.190***	0.060***
	(0.067)	(0.033)	(0.024)	(0.023)
HIGH instructions dummy	0.216	-0.208	0.984***	-0.185
	(0.739)	(0.526)	(0.365)	(0.341)
Absolute deviations × HIGH instructions dummy	0.082	-0.023	-0.096***	0.049
	(0.078)	(0.044)	(0.034)	(0.039)
Constant	1.411**	2.629***	-0.219	0.129
	(0.603)	(0.577)	(0.345)	(0.288)
Observations	765	1395	765	1395

Standard errors clustered at group level in parentheses. Includes round dummies (not reported). *, **, *** - Significant at, resp., 10%, 5% and 1%.

To examine the likelihood of being punished, Probit regressions were estimated. The dependent variable is 1 if an individual received *any* punishment in the round and 0 otherwise (Table 4, last two columns). No difference between treatments is found in the case of non-negative deviations. With negative deviations, the likelihood of punishment is increasing in the deviation of one's contribution from that of others.[19] Importantly, the HIGH instructions dummy is positive and significant, indicating that subjects are more willing to punish negative deviations in Pun-H than in Pun-L. However, the interaction term indicates that punishment is somewhat more responsive to the magnitude of the deviation in Pun-L. As probit coefficients are not readily interpretable as magnitudes, to identify the net effect on punishment likelihood, Figure 4 presents predicted probabilities of receiving punishment, as a function of negative and positive deviations.

[19] This is so also for positive deviations. This is most likely related to anti-social punishment associated with "blind" revenge (Hermann et al., 2008).

Figure 4. Predicted probabilities of receiving punishment



Vertical bars are 95% confidence intervals.

As shown in Figure 4(a), negative deviations are less likely to be punished in Pun-L than in Pun-H. The positive and significant dummy in the probit regression in Table 5 captures this. As seen, this is particularly so for small deviations – the likelihood of punishment is significantly greater in Pun-H for deviations up to 10 tokens. The negative interaction term captures this difference between treatments. Subjects in Pun-H are *less* sensitive to the size of negative deviations than are subjects in Pun-L. They are simply more likely to punish negative deviations, regardless of their magnitude. The predicted probability of receiving punishment at the average negative deviation of 6.44 tokens[20] is 0.643 in Pun-L and 0.897 in Pun-H. These are remarkably close to the actual punishment frequencies (over all sizes of negative deviations) of 0.70 and 0.92 in Pun-L and Pun-H respectively.[21] The differences are not found to be significant for non-negative deviations, i.e., for 'anti-social' punishment.

Result 5: *Subjects contributing below the mean of other group members are punished with a significantly higher frequency in Pun-H than in Pun-L.*

[20] The average negative deviation across both punishment treatments was 6.44 tokens. The average absolute negative deviation was 5.89 tokens in Pun-L and 7.59 tokens in Pun-H. The larger deviation in Pun-H does not necessarily indicate that there are more free riders in Pun-H. On the contrary, it indicates that average contribution levels were lower in Pun-L, thus necessarily making deviations also smaller in magnitude.

[21] A proportions test shows that these punishment frequencies of negative deviations are significantly different between the two treatments ($p = 0.0104$).

5. Conclusion

Carefully crafted instructions are crucial for ensuring that experimental subjects fully understand the incentive structures within the laboratory decision setting. While previous evidence suggests that the length and format of instructions can significantly affect comprehension levels of experimental subjects in public goods games, we know of no evidence that *overall behaviour* is significantly affected. We hypothesise that comprehension levels might significantly affect behaviour, and results, in more complicated settings.

In summary, in a public goods game with punishment, longer instructions with a more explicit discussion of the positive externality associated with public goods contributions are associated with more consistent targeting of low contributors and higher contributions to the public good. As discussed, studies have shown that the effectiveness of punishment in social dilemmas can be reduced in more complex decision settings. This is because the increased complexity lowers the ability of groups to coordinate on appropriate contribution norm that can then be enforced with punishment. In this sense, our study suggests that instructions that make the social dilemma more transparent to the subjects may improve the effectiveness of punishment.[22]

Our results highlight alternative instruction-formats may impact behaviour, in particular in more complex decision settings. While our results provide evidence that instruction format can affect behaviour, our LOW and HIGH instructions varied in more than one dimension. More work is thus needed to identify the particular details of experimental instructions that might affect subject behaviour.

Acknowledgements

We thank the editor Ananish Chaudhuri, Maria Bigoni, Enrique Fatas and three anonymous reviewers for helpful comments, and Sara Godoy for research assistance. Funding from EssexLab and the Universities of East Anglia and Málaga is gratefully acknowledged.

[22] While contribution behaviour is different, we do not find significant differences in earnings across treatments. See Appendix C6 for an analysis of earnings.

References

Andreoni, James (1995) "Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments", *Quarterly Journal of Economics*, 110(1), 1-21.

Chandler, Paul and John Sweller (1991) "Cognitive load theory and the format of instruction", *Cognition and Instruction*, 8(4), 293-332.

Bigoni, Maria and Davide Dragone (2012) "Effective and efficient experimental instructions", *Economics Letters*, 117(2), 460-463.

Cason, Timothy N., and Lata Gangadharan (2015) "Promoting cooperation in nonlinear social dilemmas through peer punishment", *Experimental Economics*, 18(1) 66-88.

Chaudhuri, Ananish (2011) "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature", *Experimental Economics*, 14(1), 47-83.

De Jong, Ton (2010) "Cognitive load theory, educational research, and instructional design: some food for thought" *Instructional Science*, 38(2), 105-134.

Fatas, Enrique and Guillermo Mateu (2015) "Antisocial punishment in two social dilemmas", *Frontiers in Behavioral Neuroscience*, 9, 107, 1-12.

Fehr, Ernst and Simon Gächter (2000) "Cooperation and Punishment in Public Goods Experiments", *American Economic Review*, 90(4), 980-994.

Fischbacher, Urs (2007) "z-Tree: Zurich toolbox for ready-made economic experiments", *Experimental Economics*, 10(2), 171-178.

Fischbacher, Urs, Simon Gächter and Ernst Fehr (2001) "Are people conditionally cooperative? Evidence from a public goods experiment", *Economics Letters*, 71(3), 397-404.

Gächter, Simon, Elke Renner and Martin Sefton (2008) "The Long-Run Benefits of Punishment", *Science*, 322(5907), 1510.

Herrmann, Benedikt, Christian Thöni and Simon Gächter (2008) "Antisocial punishment across societies", *Science*, 319(5868), 1362-1367.

Kirmes, Michael (2014) "Do you understand, or shall I draw you a picture? Illustrated Experimental Instructions", *Third Year Paper* submitted to the University of Arizona.

Ramalingam, Abhijit, Sara Godoy, Antonio J. Morales and James M. Walker (2016) "An individualistic approach to institution formation in public goods games", *Journal of Economic Behavior and Organization*, 129, 18-36.

Ramalingam, Abhijit, Antonio J. Morales and James M. Walker (2017), "Peer sanctioning in repeated isomorphic give and take social dilemmas", *CBESS Discussion paper 16-09.*

Rand, David G., Joseph J. Armao IV, Mayuko Nakamaru and Hisashi Ohtsuki (2010) "Anti-social punishment can prevent the co-evolution of punishment and cooperation", *Journal of Theoretical Biology,* 265(4), 624-632.

Reuben, Ernesto, and Arno Riedl (2009) "Public Goods Provision and Sanctioning in Privileged Groups", *Journal of Conflict Resolution,* 53(1), 72-93.

Reuben, Ernesto, and Arno Riedl (2013) "Enforcement of contribution norms in public goods games with heterogeneous populations", *Games and Economic Behavior,* 77(1), 122-137.

Robbett, Andrea (2016) "Sustaining cooperation in heterogeneous groups", *Journal of Economic Behavior and Organization*, 132, Part A, 121-138.

Tversky, Amos and Daniel Kahneman (1981) "The Framing of Decisions and the Psychology of Choice", *Science*, 211(4481) 453-458.