**Exploring Undermining Cooperation Effect of Punishment in Social Dilemma Contexts**

Keiko Mizuno1,2\*, Hiroshi Shimizu1

1Kwansei Gakuin University, 1-155 Uegahara 1-bancho, Nishinomiya City, Hyogo, 662-8501, Japan

2 Japan Society for the Promotion of Science, 5-3-1 Kojimachi Business Center, Chiyoda-ku, Tokyo, 102-0083, Japan

\*Keiko Mizuno (k.mizuno@kwansei.ac.jp)

Letters on Evolutionary Behavioral Science

### Method

### Participants

We determined the sample size based on a pre-test power analysis for the Mixed Model, setting the significance level α to .05, the effect size (Cohen’s d) to 0.8, the intra-class correlation (ICC) to .87, and the power to 0.8 – 0.9. A large effect size was indicated by d = 0.8, which we selected because d was 1.12 in Chen et al. (2009). We also took into account the intra-class correlations within the group, which were not reported in Chen et al. (2009). We calculated them from data from a similar public goods game conducted by Mizuno and Shimizu (2023). The sample size was determined using the formula proposed by Usami (2011). Based on these indicators, the sample size was designed to be 180-240 participants.

We recruited participants from a psychology class at a private university in Japan; they were offered course credit and Amazon gift cards as incentives for their participation. A total of 200 students participated in the study, with 58 participants identifying as male, 138 as female, 0 as other, and 3 choosing not to answer. The participants had an average age of 19.69 (SD = 2.45).

### Experimental design

The experimental design was a between-participants design with the manipulated variable, punishment, as the single factor with two levels, punished and control (unpunished).

### Procedure

Upon entering a Zoom meeting room, participants were instructed to perform the experiment in a secluded and quiet area in which they could focus. The experimenter guaranteed the anonymity of the participant's names, images, and voices. While introducing the experiment, the experimenter displayed the experimenter's video, which was turned off during the task execution. The experiment allowed a maximum of ten participants at a time, with eight participating in the primary experiment and two in a separate individual experiment (not reported here). The allocation of participants was based on their order of entry to the meeting and took into consideration the possibility of last-minute cancellations. If the number of participants was less than seven, four took part in the primary experiment, and the remaining participants engaged in individual experiments. To perform the experimental task, the participants utilized a web-based platform developed using oTree 3 (Chen, Schonger, & Wickens, 2016). The participants could select from the options presented on the screen or enter their answers by typing.

After consenting to participate in the study, the participants were informed that the task was divided into two parts, with the monetary reward depending upon their decisions in the second part, and it was emphasized that the decisions made in the task would not influence their course credits. During the first phase of the experimental task, the participants were asked to indicate their preferred allocation of funds with respect to their own share and that of others involved in the distribution for each of 52 instances. The task was specifically designed to measure both altruistic and egalitarian tendencies (Mizuno & Shimizu, 2020). During the second phase of the experimental task, the participants engaged in a repeated public goods game. Following an explanation of the rules by the experimenter, the participants were presented with a six-question quiz designed to test their understanding of the game's mechanism. Immediate feedback was given for both correct and incorrect answers. Following the quiz, the participants were requested to estimate the average number of points they believed the other participants, excluding themselves, would contribute. This measure aimed to capture their expectations regarding the cooperation of others involved in the game.

Subsequent to the public goods game, participants were requested to provide additional information, including their prior participation in similar experiments during the preceding year, familiarity with any other participants involved in the current experiment, the specific strategies they employed during the public goods game, and their individual reflections on the game. Examples of prepared reflection items included "I expected the other group members to discontinue cooperation in the absence of punishment" and "I perceived the other group members as programmed bots rather than people." Participants were also asked to provide their age and gender information, and a free-text statement form was made available for any additional comments.

Upon completion of the experiment, the participants were awarded an Amazon gift card by e-mail as compensation for their participation. The base payment was 350 yen, and additional points were earned across the total of 10 public goods games conducted. The value of each point earned was set at 1 yen, resulting in a potential reward range between a theoretical minimum of JPY 465 and a maximum of JPY 850.

### Public goods game

Participants were paired in groups of four to play a public goods game. They were informed that their group composition would remain constant throughout the experiment. During each round, each participant was given 20 points by the experimenter and decided on the number of points that they would contribute to the group, within the range of 0 to 20 points. Any uncontributed points were retained by the participants. The experimenter doubled the total points contributed by all participants and distributed them equally among the group members, with a marginal per capita return (MPCR) of 0.5. The sum of the points obtained from this distribution and the points remaining with each participant determined their total points earned. After all group members had determined their contribution, the experimenter provided feedback to the participants in the form of a table. The participant IDs (A, B, and C) remained fixed for each period, but this was not explicitly communicated to the participants. The cumulative points were not displayed, either for the participants themselves or for the other group members. The ten rounds of the public goods game were divided into two phases of five rounds each, without informing participants of the division. During the interval between the phases, the subsequent phase was explained, and participants were asked to estimate the average number of points they believed the other participants would contribute in the following rounds (belief). The total number of rounds was not disclosed.

### Conditions

In the experimental (punishment) condition, participants were informed that a punishment system would be introduced and that anyone contributing fewer than 10 points would have 12 points deducted from their contribution. In Phase 1, feedback on punishment followed the feedback on the results, and participants received feedback only for their own results. If participants contributed more than 11 points, they were told there was no punishment and that the final number of points earned in this round was XX. If the participant's contribution was less than 10 points, they were informed that 12 points were deducted, and the final number of points earned in this round was XX points. At the beginning of phase 2, the punishment system was abolished, and participants were informed that no points would be deducted regardless of the amount of contribution.

In the control condition, participants played the public goods game without punishment throughout the two phases. During instructions given between phases, participants were informed that the same task would continue in the next phase.

### Variables

**Contribution:** The amount of contribution for each participant was determined by their choice of a value from 0 to 20 points and was recorded as a single-byte integer.

**Altruism, equality:** We assessed the participants' preferences for sharing money between themselves and strangers to measure altruism and equality. We labeled the strangers as A. Each participant’s altruistic and egalitarian tendencies were estimated using the method developed by Mizuno and Shimizu (2020). The option on the left side of the display presented an equal distribution, while the option on the right side presented an unequal distribution. The money on the left side decreased in steps of ten, from 900 yen to 0 yen, as people tend to dislike unequal distributions. It was assumed that participants would initially choose the left option and that, at a certain point, their preference would change to the right option. After the left option decreased to zero, the pattern of amounts on the right side was changed. There were four different patterns for the right side options: "You get 900 yen, Mr. A gets 500 yen", "You are 700 yen, Mr. A is 500 yen", and "You are 100 yen, Mr. A is 500 yen. We presented 12 items in which left and right options were unevenly distributed to distinguish between participants who selected the left option for all choices due to complete egalitarianism and those who randomly chose the left option. The task consisted of a total of 52 questions.

**Belief:** Participants were asked to indicate their belief about other participants’ contributions by selecting from a drop-down list in response to the question, "How many points do you think, on average, other participants in the group (excluding yourself) would contribute? Please choose from 0 to 20 points."

**Experience of participating in experiments:** Participants were asked whether they had participated in a similar experiment conducted during the autumn semester of the previous year and were required to respond with a “Yes: 1” or “No: 0”.

**Acquaintance availability:** Participants were asked whether they knew anyone, such as a friend, who also participated in the experiment at the same time, with two response options: “Yes: 1” or “I don't know: 0”.

**Quiz**: Participants took a quiz on the rules of the public goods game after the game was explained to them. The quiz consisted of six questions, with responses to be made by choosing the correct number to complete the sentence.

### Additional Analyses

### Period-by-period analysis

The effects of factors were analyzed for each period. The results are shown in Table S1. The analytical model is as follows:

$$Y\_{ij}=β\_{0j}+β\_{1}X\_{j}　$$

In this equation, the variable $Y\_{ij} $represents the average of contribution, whereas $X\_{j}$ represents the treatment condition (0 for control and 1 for experimental) with $β\_{0j}$ and $β\_{1}X\_{j}$ fixed. To obtain the standardized partial regression coefficients, the objective variables were standardized.

Table S1 Period-by-period analysis



### Interaction effects of Phase and Condition

The model was further analyzed by incorporating both phase and condition interaction effects,

$Y\_{ij}=β\_{0j}+β\_{1}X\_{j}+β\_{2}phase\_{j}+β\_{3}X\_{j}\*phase\_{j}$*.*

In this equation, the variable $Y\_{ij} $represents the average of contribution, whereas $X\_{j}$ represents the average of the five contributions made in Phase 2. Besides intra-class correlations within groups, we also considered those within individuals. In this analysis, the objective variables were not standardized.

The results indicated a statistically significant interaction effect between the phase and treatment $(β\_{3}​=-3.65 (95\% CI [-4.73, -2.57]), t(197.64) =-6.60,p<.001)$. The subtest results elucidated that the Phase1 significantly augmented contributions during punishment condition $(β\_{1}=9.25 (95\% CI [6.68, 11.82]), t(119.50) =7.02,p<.001)$, while no notable difference was observed in Phase 2 $(β\_{1}=1.95 (95\% CI [-0.12, 4.02]), t(55.17) =1.84,p=.07)$.

### Regression coefficients of the full model

While the model wasn't chosen based on the BIC criteria, we have provided the regression coefficients for the full model that includes all explanatory variables,

$$\begin{array}{c}Y\_{ij}=β\_{0j}+β\_{1}X\_{j}+β\_{2}alt\_{ij}+β\_{3}equ\_{ij}+β\_{4}belief\_{ij}+β\_{5}C1\_{ij}+β\_{6}C2\_{ij}+β\_{7}C3\_{ij}　\#\left(1\right).\end{array}$$

 To obtain the standardized partial regression coefficients, the objective variables were standardized. In this equation, the variable $Y\_{ij} $represents the average of the five contributions made in Phase 2, whereas $X\_{j}$ represents the treatment condition (0 for control and 1 for experimental). Additionally, $alt\_{ij}$ represents altruism,$ equ\_{ij}$ represents equality, $belief\_{ij} $represents the expectation of cooperation from others before the start of Phase 1, $C1\_{ij}$ indicates the prior task experience, $C2\_{ij} $indicates participants acquaintance with any other participants in the experiment, and $C3\_{ij}$ represents the extent to which participants perceived the other participants as computers. Model 1 was used as the full model, with $β\_{0j}$ and $β\_{1}X\_{j}$ fixed.

　Table S2 displays the regression coefficients for each variable.

Table S2 Regression coefficients of the full model

