LEBS
Vol. 14 No. 2 (2023) 71–75
Letters on Evolutionary Behavioral Science

Exploring Undermining Cooperation Effect of Punishment in Social Dilemma Contexts

Keiko Mizuno[1,2*], Hiroshi Shimizu[1]

[1]Kwansei 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)

Previous research on social dilemmas indicates that the introduction and subsequent removal of punishment mechanisms may diminish trust among participants. Yet, empirical evidence concerning the behavioral consequences of punishment elimination on cooperation remains scarce. This study presents the concept of the Undermining Cooperation Effect of Punishment (UCEP) to operationalize the decrease in cooperation following punishment removal, in contrast to a control group never subjected to punishment. We sought to evaluate whether punishment precipitates UCEP empirically. Our pre-registered experiment yielded no significant evidence supporting UCEP; notably, cooperation levels did not endure post-punishment removal. These results contribute to the evolving discourse surrounding the intricate relationships among punishment mechanisms, trust, and cooperative behavior. They also invite further exploration into the potential counterbalancing impacts of punishment on cooperation and UCEP.

Keywords
undermining cooperation effect, punishment, public goods game

Introduction

Human beings often foster cooperation extensively, with punishment systems playing a pivotal role in this endeavor. Numerous studies have suggested that centralized punishment is a solution to secondary social dilemmas (Ozono et al., 2016; Sigmund et al., 2010). In contemporary society, law enforcement and other public authorities commonly implement punishment systems (Baldassarri & Grossman, 2011).

Conversely, implementing a centralized punishment system can yield adverse effects in various real-world contexts (Cardenas et al., 2000; Gneezy & Rustichini, 2000; Kornhauser et al., 2020). Notably, psychological research has shown that punishment can undermine trust in others (Chen et al., 2009; Irwin et al., 2014; Mulder et al., 2006). This effect may stem from shifts in the attribution of others' behavior. Without the threat of punishment, individuals often attribute others' cooperative behavior to intrinsic motivation and kindness. However, the presence of punishment can lead to attributions of cooperation to external punitive threats. Previous studies have explored this phenomenon using the Remove-the-Sanction (RTS) paradigm employed by Mulder et al. (2006). Within this paradigm, an experimental group experiences a sanction, which is later removed, while a control group never encounters a sanctioned scenario.

Although these studies consistently indicate that punishment erodes trust, findings regarding the influence of punishment on cooperative behavior remain inconsistent, with many studies not indicating significant differences. The study by Chen et al. (2009) is an exception, showing reduced cooperative behavior post-punishment removal compared to a control condition devoid of punishment; however, this study entailed both reward and punishment effects. In contrast, the pre-registered study by Mizuno and Shimizu (2023) investigated the sole impact of punishment within an RTS paradigm, revealing that post-punishment abolition cooperation levels mirrored those in the control group, exhibiting no significant variance.

The current study delineates the Undermining Cooperation Effect of Punishment (UCEP) as a decrease in cooperation within the punished group post-punishment abolition compared to a control group, as in Mizuno and Shimizu (2023). The RTS paradigm posits that punishment may reduce individuals' intrinsic trust. It is critical to demonstrate an actual decline in cooperative behavior to ascertain the detrimental impact of punishment in social dilemmas.

The differences in experimental procedures between Chen et al. (2009) and Mizuno and Shimizu (2023) could explain the occurrence of UCEP. Chen et al.'s (2009) experiment entailed two phases: participants initially engaged in five rounds of public goods games with punishment, followed by five rounds without punishment. On the other hand, Mizuno and Shimizu's experiment encompassed three phases: five rounds without punishment, five with punishment, and five rounds again without punishment. The discounting principle might explain the diverging results rooted in the structural differences of the experiments (Kelley, 1973). Both Mulder et al. (2006) and Chen et al. (2009) argue that the institution of a punishment system erodes trust within a public goods game, as participants tend to attribute others' cooperative behavior solely to the fear of punishment. In essence, participants cease cooperation because they believe that others are only cooperating due to the fear of punishment, and they would not cooperate if this punitive element were removed. In the experiment by Mizuno and Shimizu (2023), participants observed others' cooperation in the absence of punishment before its introduction,

doi: 10.5178/lebs.2023.113
Received 04 September 2023.
Accepted 16 October 2023.
Published online 30 December 2023.

potentially attributing such cooperation to innate altruism and kindness. Consequently, participants might have anticipated continued cooperation post-punishment removal, given the earlier observed cooperative behavior without punishment. In contrast, the initial introduction of punishment in Chen et al.'s (2009) experiment did not allow participants to observe cooperation in a punishment-free context, possibly leading to a stronger perception of punishment's influence on cooperative behavior compared to Mizuno and Shimizu (2023). Thus, UCEP might only manifest when punishment is established from the outset.

Based on the preceding discussion, we hypothesize that employing a two-phase structure similar to Chen et al. (2009) will reveal UCEP. This study's hypotheses do not account for the variance between states with and without punishment, focusing instead on comparing these states to a control condition. Two primary rationales underlie this analytical focus. Firstly, when contemplating implementing punishment, it's crucial to gauge its impact by comparing it to scenarios devoid of punishment (the control group). If cooperation declines post-punishment removal relative to scenarios without punishment, avoiding punishment might be prudent. Secondly, our analytical approach aims to examine the reduction in cooperation arising from decision-making shifts in repeated games. We assume that any decision-making alterations due to repetition remain consistent across both conditions. By comparing the experimental and control conditions, we aim to isolate and examine solely the UCEP, excluding the influence of repetition on decision changes.

H1: In the punishment condition, contributions following the removal of punishment will be lower in the experimental group than in the control group.

Methods

Participants

We based our sample size determination on a pre-test power analysis for a Mixed Model, setting the significance level (α) at .05, the effect size (Cohen's *d*) at 0.8, the intra-class correlation (ICC) at .87, and the power between 0.8 and 0.9. A large effect size, denoted by $d = 0.8$, was selected in light of $d = 1.12$ observed in Chen et al. (2009). We also contemplated the intra-class correlations within the group, which were not reported in Chen et al. (2009) but were derived from data from a similar public goods game conducted by Mizuno and Shimizu (2023). The sample size was determined utilizing the formula proposed by Usami (2011). Given these parameters, we aimed for a sample size between 180 and 240 participants.

We enlisted participants from psychology course students at a private university in Japan, offering course credit and Amazon gift cards as incentives ($N = 200$, 58 men and 138 women, 4 an undetermined gender). The average age of the participants was 19.69 ($SD = 2.45$).

Experimental design

The study employed a between-participants design, manipulating the single punishment factor with two levels: punished and control (unpunished).

Procedure

Participants entered a Zoom meeting room, and the experimenter instructed them to participate in the experiment in a quiet, distraction-free area. The experimenter assured the participants anonymity concerning names, images, and voices. The participants utilized a web-based platform developed using oTree 3 (Chen et al., 2016) for the experimental task.

Post-consent, participants learned that the task comprised two parts, with monetary rewards contingent on decisions made in the second part. It was emphasized that task decisions would not influence their course credits. In the first phase of the experimental task, the participants indicated their preferred fund allocation regarding their share and others' shares across 52 instances, designed to gauge altruistic and egalitarian tendencies (Mizuno & Shimizu, 2020). Subsequently, participants took part in a repeated public goods game. Following the experimenter's explanation of the rules, a six-question quiz tested their understanding of the game's mechanics, with immediate feedback for correct and incorrect answers. Post-quiz, participants estimated the average points they anticipated other participants would contribute. This measure aimed to capture the participants' expectations regarding others' cooperation in the game.

Following the public goods game, participants provided additional data, including prior participation in similar experiments during the preceding year, familiarity with other participants in the current study, strategies employed during the public goods game, reflections on the game, age, and gender. Reflection items such as "I expected other group members to discontinue cooperation in the absence of punishment" and "I perceived the other group members as programmed bots rather than people" were included. A free-text statement form was also available for additional comments.

Upon completing the experiment, participants received an Amazon gift card via email as compensation, with a base payment of 350 yen and additional points earned across the ten public goods games conducted. Each point earned equated to 1 yen, yielding a potential reward range between a theoretical minimum of JPY 465 and a maximum of JPY 850.

Public goods game

Participants formed groups of four to engage in a public goods game, with consistent group composition maintained throughout the experiment. During each round, participants received 20 points from the experimenter and decided on the number of points they would contribute to the group (ranging from 0 to 20 points). The participants retained the uncontributed points. The experimenter doubled the total points contributed by all participants and equally distributed them among group members, with a marginal per capita return of 0.5. The sum of the points from this distribution and the remaining points determined each participant's total points earned. The experimenter provided feedback in table format after all the group members had determined their contribution. Participant IDs (A, B, and C) remained fixed across each period, albeit without explicit communication to participants. Cumulative points were not displayed, neither to the participants nor to the other group members. The ten rounds of the public goods game were split into two

phases of five rounds each, with participants uninformed of this division. Between phases, the upcoming phase was explained, and participants were prompted to estimate the average points they believed other participants would contribute in the subsequent rounds (belief). The total number of rounds was undisclosed.

Conditions

In the experimental (punishment) condition, participants were informed of a punishment system in which contributions below 10 points would incur a 12-point deduction. In Phase 1, feedback on punishment followed the feedback on the results, with participants only receiving feedback for their own results. If contributions exceeded 11 points, participants were informed there was no punishment and the final points earned in that round. If contributions were below 10 points, participants were informed of a 12-point, along with the final points earned in that round. As Phase 2 commenced, the punishment system was abolished, with participants informed that no points would be deducted, irrespective of the contribution amount. In the control condition, participants played the public goods game without punishment across both phases. The instructions between the two phases informed the participants that the same task would continue in the next phase.

Analytical design

Utilizing R version 4.2.1 (R Core Team, 2019), we examined the effects of control and experimental factors using a mixed model, assuming a variable effect on the intercept for a four-person population. The model is outlined below:

$$Y_{ij} = \beta_{0j} + \beta_1 X_j + \beta_2 \mathrm{alt}_{ij} + \beta_3 \mathrm{equ}_{ij} + \beta_4 \mathrm{belief}_{ij} + \beta_5 C1_{ij} + \beta_6 C2_{ij} + \beta_7 C3_{ij} \quad (1).$$

In this equation, i represents an individual and j stands for a group. 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 $\beta_1 X_j$ fixed. The remaining control variables, including $\beta_2 alt_{ij}$, were selected and forced to minimize the Bayesian Information Criterion (BIC). Model selection was conducted via maximum likelihood using the lmer function from the lmerTest package (Kuznetsova et al., 2017), while the assessment of experimental effects was conducted through restricted maximum likelihood. The hypothesis would be considered supported if the effect β_1 of X_j on Y_{ij} was negative and significant in a two-tailed test at the 5% level for models with the lowest BIC.

Registration

The hypotheses, methods, analytical design, and R code for analysis pertinent to this study were pre-registered on the Open Science Framework (OSF) before the commencement of the experiment. The pre-registration for this study is accessible at https://osf.io/2pts5. While the original pre-registration is in Japanese, an English version is available at https://osf.io/8nxsv/files/osfstorage/648c6ea56c0981005c7d245a.

Results

Figure 1 illustrates the contribution trend across ten periods. In the experimental (punishment) condition, the average contribution was 14.12 (SD = 3.36) during Phase 1 and 7.67 (SD = 6.29) during Phase 2. Conversely, in the control condition, the average contribution was 8.51 (SD = 5.28) during Phase 1 and 5.72 (SD = 4.86) during Phase 2. The 95% confidence intervals for the average contribution during Period 6, which directly followed the cessation of punishment, spanned [8.78, 11.32] for the punishment and [7.03, 8.26] for the control condition.

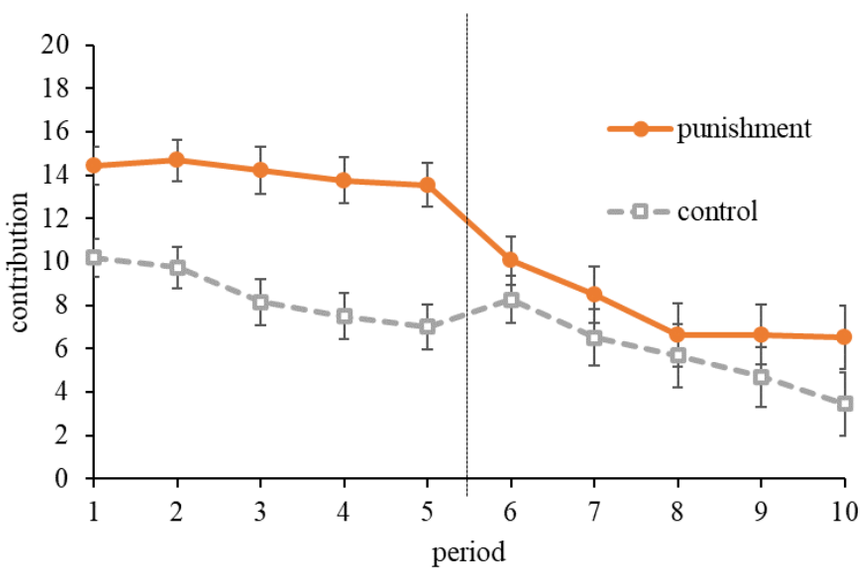

Figure 1. Contributions per condition.

Based on BIC, we selected the β_{oj} and $\beta_1 X_j$ model. We scrutinized the contribution amounts using these as the objective variable to validate the elevated contribution in the punishment condition during Phase 1. The results indicated a significant effect of the treatment (β_1 = 1.07 (95% CI [0.74,1.40]), $t(48.00) = 6.32$, $p < .001$). We executed a linear mixed model to evaluate the hypothesis that contributions after removing punishment would be lower in the experimental group relative to the control group, as delineated previously in the Analytical Design section. We standardized the dependent variable to derive a standardized partial regression coefficient. The main effect of factor X_j was not significant (β_1 = 0.34(95% CI[−0.09,0.78]),$t(48.00) = 1.54$, $p = .130$), indicating that the hypothesis was not supported.

Discussion

This study delved into the potential occurrence of a UCEP, extending the insights from Chen et al. (2009). We hypothesized that following the removal of punishment, the punishment condition would yield lower cooperation levels than the control condition. However, the results did not validate this hypothesis, as we observed no adverse effects of punishment.

In this investigation, we characterized UCEP as the phenomenon where, following the initiation and subsequent withdrawal of punishment, the contribution from the experimental group is lower than that from the control group who experienced no punishment. We explored three potential scenarios to comprehend the absence of UCEP. Initially, we considered the likelihood of a floor effect, where the control group's contribution might have been so minimal that detecting UCEP became challenging. Yet, glancing at the error bars in Figure 1 dissuades the possibility of a floor effect.

Furthermore, we speculated whether the cooperation level in our control group was lower than those documented in previous studies, which could have facilitated the detection of UCEP if the control group had exhibited heightened cooperation alongside the punishment withdrawal. A review of prior research to discern if our control group's contribution was unusually low indicated varying levels of cooperation in control groups of earlier studies. For instance, in the peer punishment experiment by Fehr and Gächter (2002), the control group contributed around 35% of the endowment. In studies with automatic punishment systems resembling ours, the control group's cooperation levels were 56% in Mulder et al. (2006) Study 1, 53% in Chen et al. (2009) Study 1, 63% in Irwin et al. (2014), and 38% in Mizuno and Shimizu (2023). Notably, the design of some studies, like Mulder et al. (2006) and Irwin et al. (2014), differed from ours as they encompassed a single period with and without punishment, making direct comparisons with our study, which had five periods challenging. In our study, the average contribution from the control group in the sixth period (immediately after ceasing punishment) constituted approximately 39% of the endowment. Compared to the 28% observed in our study, other research, especially Chen et al. (2009), which identified UCEP and recorded a 50% contribution, exhibited higher rates. This discrepancy suggests that the lower contributions from our control group might underlie our inability to detect UCEP. Future inquiries may probe whether fostering higher cooperation by initially communicating the benefits of cooperation to both groups with the message, "it's beneficial to cooperate," unveils UCEP.

Lastly, the failure to observe UCEP might stem from counterbalancing forces. These forces separate into two categories: those that uphold cooperation (and hence, do not decline throughout the period)—even after punishment withdrawal—and those that substantially reduce cooperation after ceasing punishment. A plausible factor perpetuating post punishment cooperation could be the formation of societal norms. For example, studies by Mulder et al. (2009) and Nolan (2017) illustrate how punishment fortifies the norm against noncooperation. Given such findings, it is conceivable that once individuals or groups internalize these norms, cooperation might endure even after punishment is removed. These counterbalancing forces, manifesting at both individual and group levels, warrant exploration in subsequent research to unravel this hypothesis further.

Acknowledgments

This research received funding from the Japan Society for the Promotion of Science (JSPS) through its Grants-in-Aid for Scientific Research program, under grant numbers 21J22689 and 22KJ3053.

Author contribution

KM designed the research plan, collected and analyzed the data, and wrote the paper. SH provided valuable guidance on research design, data collection, and analysis, and reviewed the manuscript of the paper.

Ethical statement

All participants provided informed consent in adherence to the Kwansei Gakuin University Regulations for Behavioral Research with Human Participants. The Kwansei Gakuin University Institutional Review Board for Behavioral Research with Human Participants approved this experiment. We conducted all methods in strict compliance with the pertinent guidelines and regulations.

Data accessibility

The datasets pertinent to this study are accessible at https://osf.io/8nxsv/.

Supplementary material

Electronic supplementary material is available online.

References

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11023–11027. https://doi.org/10.1073/pnas.1105456108

Cardenas, J. C., Stranlund, J., & Willis, C. (2000). Local environmental control and institutional crowding-out.

World Development, *28*(10), 1719–1733. https://doi.org/10.1016/S0305-750X(00)00055-3

Chen, X. -P., Pillutla, M. M., & Yao, X. (2009). Unintended consequences of cooperation inducing and maintaining mechanisms in public goods dilemmas: Sanctions and moral appeals. *Group Processes and Intergroup Relations*, *12*(2), 241–255. https://doi.org/10.1177/1368430208098783

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies*, *29*(1), 1–17. https://www.jstor.org/stable/10.1086/468061

Irwin, K., Mulder, L., & Simpson, B. (2014). The detrimental effects of sanctions on intragroup trust: Comparing punishments and rewards. *Social Psychology Quarterly*, *77*(3), 253–272. https://doi.org/10.1177/0190272513518803

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128. https://doi.org/10.1037/h0034225

Kornhauser, L., Lu, Y., & Tontrup, S. (2020). Testing a fine is a price in the lab. *International Review of Law and Economics*, *63*, Article 105931. https://doi.org/10.1016/j.irle.2020.105931

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Mizuno, K., & Shimizu, H. (2020). Fubyōdō kihi moderu no keikenteki kentō: Beizu tōkei moderingu ni yoru moderu hikaku [An empirical study of the inequality aversion model: Model comparison using Bayesian statistical modeling]. *KG Sociological Review*, *9*, 41–51.

Mizuno, K., & Shimizu, H. (2023). Shakaiteki jirenma jōkyō ni okeru batsu no gyakukōka no kentō: "Remove the sanction" paradaimu o mochiite [Examining the detrimental effects of punishment in social dilemma situations in the context of the "remove the sanction" paradigm]. *Japanese Journal of Social Psychology*, *38*(3), 51–58. https://doi.org/10.14966/jssp.2205

Mulder, L. B., van Dijk, E., De Cremer, D., & Wilke, H. A. M. (2006). Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology*, *42*(2), 147–162. https://doi.org/10.1016/j.jesp.2005.03.002

Mulder, L. B., Verboon, P., & De Cremer, D. (2009). Sanctions and moral judgments: The moderating effect of sanction severity and trust in authorities. *European Journal of Social Psychology*, *39*(2), 255–269. https://doi.org/10.1002/ejsp.506

Nolan, J. M. (2017). Environmental policies can buttress conservation norms. *Society and Natural Resources*, *30*(2), 228–244. https://doi.org/10.1080/08941920.2016.1209266

Ozono, H., Jin, N., Watabe, M., & Shimizu, K. (2016). Solving the second-order free rider problem in a public goods game: An experiment using a leader support system. *Scientific Reports*, *6*, Article 38349. https://doi.org/10.1038/srep38349

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sigmund, K., De Silva, H., Traulsen, A., & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, *466*(7308), 861–863. https://doi.org/10.1038/nature09203

Usami, K. (2011). Kaisōteki na dēta shūshū dezain ni okeru nigun no heikinchisa no kentei, suitei no tameno sanpuru saizu ketteihō to sūhyō no sakusei [A unified method for determining the sample size needed for evaluation of mean differences in hierarchical research designs]. *Japanese Journal of Educational Psychology*, *59*(4), 385–401. https://doi.org/10.5926/jjep.59.385