The Punisher's Dilemma: Uncertainty and the Coexistence of Motivations in Social Dilemmas*

Hyoji Kwon[a] **Yukihiko Funaki**[b]

Abstract

This study uses uncertainty in the payoff function of a public goods game to distinguish between different motivations for imposing costly punishments in social dilemma situations, specifically highlighting the coexistence of these motivations. There are two possible motivations: reciprocity and inequality aversion. By introducing uncertainty, participants are prevented from predicting others' contributions based solely on their payoffs. In this situation, participants must choose between others' contributions and others' payoffs as the criteria for punishment. Our results reveal heterogeneity in punishment motivations, leading to the identification of different types of punishers: the self-interested type, the reciprocal type, the inequality-averse type, and the "other" type, who exhibits inconsistency. Additionally, the reciprocal type strongly punishes free-riding behaviors while also imposing some punishment for payoff inequality. These findings highlight that inequality aversion is a critical motivation for punishment—some individuals rely solely on inequality aversion, while others incorporate it into their punishment based on reciprocity. Notably, payoff inequality appears to play a crucial role in motivating punishment under uncertainty, regardless of the norm of cooperation.

Keywords: Costly punishments, Uncertainty, Reciprocity, Inequality aversion, Public goods game
JEL classification: C91, D63, D91

* No conflict of interest: "This paper has no conflicts of interest to disclose."
[a] University of Hyogo, kwon@em.u-hyogo.ac.jp
[b] Waseda University, funaki@waseda.jp

1. Introduction

People punish free-riding and other undesirable behaviors. People sometimes still tend to punish even if there are no future benefits, only costs. However, punishment of free riders, which also encourages cooperation, is frequently observed (Balliet et al., 2011). What, then, makes people punish others even when it is hard to expect any future benefit?

This study uses uncertainty in the payoff function of a public goods game to distinguish between different motivations for imposing costly punishments in social dilemma situations, specifically highlighting the coexistence of these motivations. When someone free rides, the action has two undesired components. One is the betrayal behavior, which is unkind and selfish, and the other is the inequality between the free rider's higher payoff and the cooperator's lower payoff. Punishment is a reaction to undesired things, and there are two streams of research about the motivations behind punishment. The first motivation is reciprocity, which leads to retaliation against undesired behaviors, and the second is inequality aversion, which reduces undesired inequality. These two motivations have been studied for a long time; however, the streams of research on costly punishments in social dilemmas consider only whether two or more motivations are able to coexist within an individual, and only a few studies have considered both motivations concurrently.

We believe that there is also heterogeneity in the motivation to punish in public goods games, and we try to classify subjects according to their motivations. In our experiment, to classify subjects according to their motivation for punishment and to verify the characteristics of the subjects exhibiting each motivation, we conduct modified public goods games with costly punishments. Since reciprocity is based on others' behaviors and inequality aversion on unequal payoffs, we use an uncertainty factor as noise, which weakens the link between behaviors and payoffs. This noise is randomly and individually added to the payoffs in a public goods game after the contribution decisions are made. For example, when negative noise is added to a free rider's payoffs in a public goods game, it decreases the free rider's high payoffs. Thus, people can choose who should be punished: free riders, high earners, or both. Based on the individuals' punishment patterns in these games, we estimate and classify the subjects' types. Then, we analyze each type's behavioral pattern in more detail.

In our results, the share of the reciprocal type is the highest, but a significant percentage of participants punish solely based on inequality aversion. As a robustness test, we conduct random income games (Dawes et al., 2007)[1] and explore whether intentions matter only to the reciprocal type. As expected, participants classified as reciprocal type exhibit the greatest decrease in punishments. This result clarifies our classification by highlighting the distinction between the reciprocal type and the inequality-averse type when there is no room for reciprocity. Moreover,

[1] The procedure for the random income game in our experiments is explained in Section 2.

we find that the reciprocal type strongly punishes free-riding behaviors but also imposes some punishment for payoff inequality. In contrast, participants who do not punish at all in the baseline game are the least cooperative. Additionally, we observe differences by gender and social characteristics: female participants are more likely to belong to the inequality-averse type, while male participants are more likely to be classified as the self-interested type. Inequality-averse participants are generally less trusting but score higher in social acceptance. These findings suggest that under conditions of uncertainty, inequality aversion emerges as a crucial motivation, and at times, achieving equality in outcomes becomes more important than maintaining the norm of cooperation, creating a punisher's dilemma between fairness and cooperation.

2. Experimental Design

Our experiments are public goods games with costly punishments Each session is composed of two games. All subjects participate in both games in order, and we call the first game the baseline game and the second game the random income game (Dawes et al., 2007). After seating all participants, we distribute the instructions for the baseline game only because the random income game starts suddenly.

At the beginning of each round in the baseline game, all participants are endowed with 20 tokens and divided into new four-member groups. Since the group composition changes randomly every round, there is little room to consider the future benefits from punishing others. Each round consists of two stages: the contribution stage (Stage 1) and the punishment stage (Stage 2). In Stage 1, participants decide the amount x_i to contribute to a public account, where $0 \leq x_i \leq 20$ and the marginal per capita return on the public account is 0.5. To distinguish between the motivations for punishment, we add the uncertainty factor ε_i to the Stage 1 payoff function. Hence, i's payoff at the end of Stage 1 is given by:

$$\pi_i = 20 - x_i + 0.5 \sum_{j=1}^{4} x_j + \varepsilon_i,$$

where ε_i is an integer in [−8, 8] and follows a random sequence with mean 0 over all rounds. The sequence of ε_i follows a uniform distribution, and the subjects do not know this distribution. However, they do know that ε_i is assigned randomly. This uncertainty factor weakens the causal relation between behaviors and payoffs. Therefore, when subjects want to punish others, they must choose their criterion, behavior or payoff. This allows us to classify the subjects according to their motivations. All participants have their own sequence of ε_i, so each group member may be assigned a different ε_i.

When all participants decide their contributions to the public good, they move on to Stage 2 and receive nine additional tokens that can be used to inflict punishments. In this stage, the participants are informed of the other three members' contributions and payoffs from the first

stage. Of course, participants know their own contributions and payoffs. They can punish the other members by using their tokens—up to three for each member—and each token decreases the target's payoff by three tokens. Hence, subject i's final payoff for each round is as follows:

$$\pi_i = 20 - x_i + 0.5\sum_{j=1}^{4} x_j + \varepsilon_i + 9 - \sum_{j\neq i}^{4} p_{ij} - 3\sum_{j\neq i}^{4} p_{ji}.$$

The baseline game has ten rounds, and participants know these payoff functions and processes, including the number of rounds.

After finishing the tenth round of the baseline game, we ask participants to play an additional game, a random income game (Dawes et al., 2007). In the random income game, the contribution stage is skipped. The reason why the contribution decision is eliminated is to clarify the subjects' motivations as follows: people who punish others for their contributions will stop punishing in a random income game because there are no intentional behaviors, and those who punish others for their payoffs will continue to impose punishments. The random income game is played for a total of three rounds, although participants are not aware of the total number of rounds that will be played in order to avoid end-game effects. The random income game is the same as the baseline game except during Stage 1. Each participant makes only a punishment decision based on the others' computer-determined contributions. For a more rigorous comparison of the two treatments, we use the data from three rounds in the baseline game. We distribute new instruction after finishing the baseline game and the computer reads it aloud. Through this random income game, our classification of motivation types based on the punishment behaviors in the baseline games will be verified.

Additionally, both before starting the baseline game and after finishing the random income game, we ask the following ex ante and ex post questions: "If you decide to reduce others' payoffs, what information do you think you would need?" and "What information do you think has actually affected your decision to reduce others' payoffs thus far?". Participants are asked to choose one of the following two options: 1) others' contribution to the public goods and 2) others' payoffs after the contribution decision has been made. From the answers to these questions, we can make conjectures about the relation between revealed motivations and stated motivations. Figure 1 briefly shows the overall experimental procedure.

Finally, to identify the characteristics of each type, we distribute a questionnaire.

The experiments were conducted at Waseda University during Summer and Fall 2018. We ran five sessions with 24 to 28 subjects in each session, and all 136 participants were students from Waseda University. The experiments were computerized with z-Tree (Fischbacher, 2007). Each session lasted 60 minutes, and the show-up payment was ¥700 (≈ $6.2), and average earnings were ¥1890 (≈ $16.6).

3. Results

3.1 Estimation of Subject Types

In our experiments, subjects face the problem of what to punish: others' behaviors or their payoffs. Thus, we hypothesize that subject types can be distinguished based on the motivations for punishment: 1) The self-interested type (Type S), 2) The reciprocal type (Type R), 3) The inequality-averse type (Type IA), and 4) The "other" type (Type O).

Carpenter and Matthews (2012) shows that subjects punish based on two norms: 1) a comparison of the individual contributions of the subject himself or herself and the target and 2) a comparison of the group average contribution and that of the target. Thus, we consider two regression models corresponding to individual and group comparisons.

To classify subject types based on this hypothesis, we have the following two models:

$$p_{ij} = \beta_1 + \beta_2 \min\{Con_j - Con_i, 0\} + \beta_3 \max\{Pay_j - Pay_i, 0\} \qquad (M1)$$

$$p_{ij} = \beta_4 + \beta_5 \min\{Con_j - GCon_{-j}, 0\} + \beta_6 \max\{Pay_j - GPay_{-j}, 0\} \qquad (M2)$$

where pij on the left-hand side is the number of punishment tokens that i spends on target j. In M1, Con_j is target j's contribution, Con_i is i's own contribution, Pay_j is target j's payoff in Stage 1, and Pay_i is i's own payoff in Stage 1. In M2, $GCon_{-j}$ is the group average contribution excluding that of target j, and $GPay_{-j}$ is the group average payoff excluding that of target j. Because of multicollinearity, we separate the variables in these two models according to two criteria: deviations from one's own contribution/payoff and deviations from the group average.

We determine the types of subjects through the following process: 1) We select coefficients with high significance levels with the lowest p value in both Models M1 and M2. 2) When two or more coefficients have the same significance level of $p = 0.0000$ in both models, we compare their magnitudes. We normalize all variables ($\min\{Con_j - Con_i, 0\}$, $\min\{Con_j - GCon_{-j}, 0\}$, $\max\{Pay_j - Pay_i, 0\}$ and $\max\{Pay_j - GPay_{-j}, 0\}$) to a range of 0 to 1 before comparing the magnitudes of the coefficients. The results of the type classification are shown in Table 1.

Type	Number of subjects	Percentage
Self-interested (Type S)	52	38%
Reciprocal (Type R)	40	29%
Inequality Averse (Type IA)	35	26%
Other (Type O)	9	7%
Total	136	100%

Table 1 Type Classification.

3.2 Main Results

First, Type R is the most cooperative, and Type S is the least cooperative. Figure 1 shows the average contributions by round for each type. This result provide evidence for our classification of types. Individuals classified as Type R, who punish others' undesired behaviors, appear to focus on others' behavior because they behave cooperatively. On the other hand, Type S subjects, who do not punish at all, are the most uncooperative.

Second, Type R reduce their punishment significantly when there is no intention behind others'

behaviors. On the other hand, Type IA also tend to decrease their punishment, but they maintain their punishment for payoff inequality regardless of others' intentions. Table 2 presents the results of linear regressions that uses the baseline game and the random income game. The regression models are as follows:

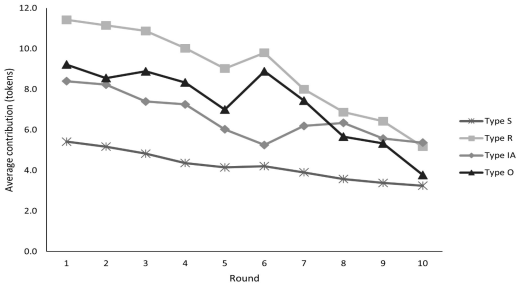

Figure 1 Cooperative Behaviors by Type.

$$p_{ij} = \beta_1 + \beta_2 Tr + \beta_3 min\{Con_j - Con_i, 0\} + \beta_4 max\{Pay_j - Pay_i, 0\} + \beta_5 Tr \times min\{Con_j - Con_i, 0\} + \beta_6 Tr \times max\{Pay_j - Pay_i, 0\} \quad (M3)$$

$$p_{ij} = \beta_7 + \beta_8 Tr + \beta_9 min\{Con_j - GCon_{-j}, 0\} + \beta_{10} max\{Pay_j - GPay_{-j}, 0\} + \beta_{11} Tr \times min\{Con_j - GCon_{-j}, 0\} + \beta_{12} Tr \times max\{Pay_j - GPay_{-j}, 0\} \quad (M4)$$

where Tr is a dummy variable that equals 0 for the baseline game and 1 for the random income game. We include interaction variables between the treatment and the motivations. $Tr \times min\{Con_j - Con_i, 0\}$ and $Tr \times min\{Con_j - GCon_{-j}, 0\}$ indicate the interaction between Tr and reciprocity, and $Tr \times max\{Pay_j - Pay_i, 0\}$ and $Tr \times max\{Pay_j - GPay_{-j}, 0\}$ indicate the interaction between Tr and inequality aversion.

Dependent variable: Punishment level	Type R	Type IA	Type O
M3 Individual comparison			
Tr (0: baseline, 1: RI)	0.0083 (0.6125)	−1.2995** (0.4713)	−6.0141* (2.7622)
$\min\{Con_j - Con_i, 0\}$	−0.4361*** (0.0675)	0.0549 (0.0625)	0.2639 (0.2465)
$\max\{Pay_j - Pay_i, 0\}$	0.1194* (0.0480)	0.1988*** (0.0480)	0.1743 (0.1474)
$Tr \times \min\{Con_j - Con_i, 0\}$	0.3165*** (0.3165)	0.0563 (0.0983)	0.9972 (0.7561)
$Tr \times \max\{Pay_j - Pay_i, 0\}$	−0.0197 (0.0690)	0.0185 (0.0715)	0.6565 (0.3826)
Constant	−4.6678*** (0.6389)	−2.2559*** (0.3855)	−3.7698* (1.4502)
M4 Group comparison			
Tr (0: baseline, 1: RI)	−1.0897 (0.5972)	−1.7651*** (0.4876)	−4.4788* (2.1286)
$\min\{Con_j - GCon_{-j}, 0\}$	−0.5005*** (0.0797)	−0.0013 (0.0734)	0.1431 (0.3218)
$\max\{Pay_j - GPay_{-j}, 0\}$	0.1155* (0.0547)	0.2701*** (0.0512)	0.1813 (0.2038)
$Tr \times \min\{Con_j - GCon_{-j}, 0\}$	0.3012** (0.1102)	−0.0894 (0.1120)	-0.1563 (0.5429)
$Tr \times \max\{Pay_j - GPay_{-j}, 0\}$	0.0465 (0.0824)	0.0087 (0.0721)	0.1067 (0.3360)
Constant	−4.0177*** (0.5615)	−2.4006*** (0.3729)	−4.3030** (1.6100)
Number of observation	588	546	150

Table 2 Results of Tobit Regressions.

Third, there are gender differences among the types. Figure 2 describes the gender distribution by type. This result is in line with results from studies of gender differences in fairness considerations; namely, women usually prefer fairness, and men are more likely to favor efficiency.

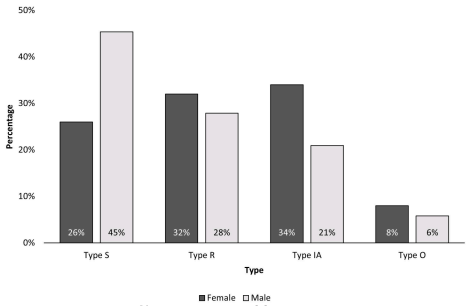

Figure 2 Gender Differences.

References

Balliet, D., Mulder, L.B., Van Lange, P.A. (2011). Reward, punishment, and cooperation: a meta-analysis. Psychological Bulletin, 137(4), 594.

Carpenter, J.P., & Matthews, P.H. (2012). Norm enforcement: anger, indignation, or reciprocity? Journal of the European Economic Associ- ation, 10(3), 555–572

Dawes, C.T., J.H. Fowler, T. Johnson, R. McElreath, and O. Smirnov. 2007. Egalitarian motives in humans. *Nature* 446.

Fischbacher, U. 2007. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10 (2): 171–178 .