# An Effect of Nudges by Moral Messages on Third-Party Punishment §1

Tetsuo Yamamori[a] Kazuyuki Iwata[b]

## Abstract

This study experimentally explores how norm-nudging, which simply tells or reminds people about the right thing to do, without stating empirical and normative information, influences people's social preferences and willingness to punish as third parties in a simple public goods game. We found that a simple moral message that reminds subjects of altruism, such that they should consider the payoffs of other subjects, enhances subjects' reciprocity and willingness to impose punishment as third parties on those who make unfair investments.

Keywords: nudge; public goods game; third party punishment; economic experiment
JEL Classification: C92, D91, H42

## 1. Introduction

During the Covid-19 pandemic, governments and local administrations worldwide had developed and issued guidelines for new lifestyles to prevent the spread of infection and had implemented various interventions encouraging compliance with such guidelines. In particular, in countries such as Japan, where domestic laws make it difficult to enforce compulsory interventions such as urban lockdowns, various slogans were devised to request or instruct people to engage in voluntary infection-preventive behavior. Examples of such slogans include "stay at home," "avoid the three Cs (closed spaces, clouded spaces, and close contact)," and "compassion vaccine," which were usually accompanied by altruistic messages such as "in order to save the lives of people close to you." The above slogans for infection-preventive behavior were intended to tell or remind people about the right thing to do by attaching moral messages.

These interventions were a variant of "norm-nudge" that attempts to induce people into socially desirable behavior by eliciting or changing existing social norms through the manipulation of social expectations (Bicchieri and Dimant 2022). This is typically done by directly stating what others do (empirical information) and/or approve (normative information), or by presenting messages about the right thing to do. In recent years, much attention has been devoted to norm-nudge for promoting pro-social behavior in situations involving conflict between individual interests and social benefits, such as social dilemmas.

§1 This research was financially supported by a grant-in-aid provided by Dokkyo University.
a Corresponding author: Faculty of Economics, Dokkyo University, 1-1, Gakuen-cho, Soka, Saitama 340-0042, Japan. Phone: +81-48-943-2217, Fax: +81-48-943-2217, Email: yamamori@dokkyo.ac.jp.
b Faculty of Economics, Matsuyama University, 4-2 Bunkyo-cho, Matsuyama, Ehime 790-8578, Japan. Email: iwata.kazuyu@gmail.com

Considerable field research has found effective messages in situations involving taxation (e.g., Hallsworth et al. 2016), electricity usage (e.g., Allcott 2011), water conservation (e.g., Ferraro and Price 2013), littering (e.g., Cialdini et al. 1990), and preventive behavior against the spread of an infection (e.g., Sasaki et al. 2021).

This study contributes to the understanding of the effectiveness of norm nudging by reinforcing altruistic punishment. Altruistic punishment is punishment by individuals (usually third parties), which is costly for punishers but does not result in any material gain to them. Social norms are enforced because of the expectation that violations of behavioral standards will be punished (Fehr and Fischbacher 2004). Therefore, a norm-nudge that intensifies altruistic punishment promotes people to engage in more pro-social behavior.

As a norm-nudge, we especially focus on moral messages that simply tell or remind people about the right thing to do without stating empirical and normative information (hereafter, we refer to such a message as a simple moral message). For policymakers, norm-nudging with a simple moral message is more convenient to use than the other norm-nudge relying on empirical and normative information because sending a simple moral message does not require any knowledge of such information. However, the empirical effectiveness of simple moral messages has been inconsistent. For example, Cialdini et al. (1990) reported that a simple message such as "Please do not litter" had a great effect on littering behavior in their experiment, wherein they manipulated the salience of normative messages. However, Bicchieri and Dimant (2022) pointed out that such a simple moral message can backfire because it may provide wrong type of information. That is, the fact that there is a need for policymakers to tell people the right things to do implies that many people behave badly, which in turn means that people misbehave because they are less likely to face social sanctions. Therefore, it is important to examine whether a simple moral message enhances altruistic punishment.

To explore how simple moral messages affect people's willingness to engage in altruistic punishment, we conducted a laboratory experiment using a linear public goods game with third-party punishment. There were four players, two of which (investors) played the public goods game, and the others (third parties) had the opportunity to punish the investors according to their contributions. Our experiments consisted of two decision stages, and all the subjects were assigned the roles of both an investor and a third party as follows. In Stage 1, all the subjects were divided into groups consisting of two subjects, each of whom received 100 points (experimental currency unit) and had to choose points of investment in the joint account (public goods) that they shared with the other subject in the same group. The sum of all points in the joint account was multiplied by 1.4 and divided equally between the two investors. In Stage 2, all the subjects received an additional 20 points, regardless of the results in Stage 1; as a third party, they could use these points to reduce the payoffs earned in Stage 1 by another subject in a different group. If a third party reduced some points from the payoffs of an investor in a different group, they had to pay $1/4^{th}$ of the reduced points from their additional points, whereas the points they did not use were their payoffs in Stage 2.

Our experimental design comprised two treatments: baseline (without moral messages) and nudge (with moral messages). Both the treatments were essentially the same, except that the message "You should consider not only your profits but also the payoffs of your partner and try to invest a significant number of points into the joint account" was written as an instruction and displayed on the computer screen in Stage 1 in the nudge treatment.

The purpose of this study is to explore how this moral message influences the social preferences of the subjects on the payoff distributions of two investors and the subjects' willingness to punish the investors as a third party. Therefore, regardless of the treatments, we adopt a variant of the strategy methods in both Stages 1 and 2 to elicit the subjects' preferences for payoff distributions in the public goods game and for

conditional punishment as a third party. In Stage 1, we adopt the following strategy method similar to Fischbacher et al. (2001): Each subject had to make two types of decisions, one was regarding "unconditional investment" and the other was regarding "conditional investments." The subjects had to make both types of decisions without knowing the decisions of the others. When it came to the "unconditional investment," each subject had to decide how many points they wanted to invest into the joint account from 0 to 100 points with an increment of 20 points (i.e., six options in total). When it came to the "conditional investment," each subject had to decide conditional investment from the same six options for each possible investment that their partner could choose as their "unconditional investment." After the end of Stage 1, the computer randomly selected one of the two investors in the group. The "unconditional investment" was adopted as the actual investment of the one investor selected by the computer, and the "conditional investment" was adopted as the actual investment of the other.

To detect the conditional punishment of the subjects, we also adopted the following strategy method in Stage 2. For each group, two subjects from the other groups were assigned to third parties (another group). Each third party could punish a different investor in this group and the investor who each third party could punish was determined randomly in advance. A third party had to decide the number of points to be reduced from the payoffs earned in Stage 1 by the investor for some possible combinations of investments of the two investors in the group without knowing their realized investment amounts. While the possible combinations of investment amounts in Stage 1 were 36 ($=6 \times 6$) patterns in total, we assigned each subject only 12 patterns, including the realized one to mitigate their workload.

Our main findings are as follows. First, investors' conditional investment is upward sloping: their investment points increase as the pair's investment increases. Moreover, this tendency was significantly stronger in the nudge treatment than in the baseline treatment. In other words, the simple moral message has enhanced the subjects' reciprocity. Second, the greater the difference in investment amounts between investors, the greater is the punishment by a third party for those who invested less. Moreover, this tendency was significantly stronger in the nudge treatment than in the baseline treatment. That is, the simple moral message has enhanced the subjects' willingness to punish those who make unfair investments.

The remainder of this paper is organized as follows: Section 2 describes our experimental design. Section 3 presents the experimental results. Section 4 summarizes the results and discusses their implications.

## 2. Experimental design and procedure

### 2.1. Underlying game

The underlying game in our experiment is the following linear public goods game with third-party punishment. There are four players, two of which (investors) play the public goods game and the others (third parties) have the opportunity to punish the investors according to the results of the public goods game. In Stage 1, the investors receive an endowment of 100 points (experimental currency unit) in each of their personal accounts and have to decide how many of those points to invest in a joint account (the public goods) shared with the other investor without knowing the other investor's decision. Every point in the personal account increases individual earnings by one point. The sum of all points in the joint account is multiplied by 1.4 and divided equally between the two investors. Thus, the marginal payoff of a contribution to the joint account is 0.7 points and the payoffs of each investor $i \in \{1, 2\}$ in Stage 1 is given by:

$$\pi_i^1 = 100 - g_i + 0.7(g_1 + g_2), \tag{1}$$

where $g_i$ is the number of points that investor $i$ invests in the joint account.

In Stage 2, the third parties have the opportunity to punish the investors based on their investments in Stage 1. Each third party receives an endowment of 20 points and can use these 20 points to reduce the payoffs of one predetermined investor. One investor can be punished by one third party, and the other investor can be punished by another third party. Let $k(i) \in \{3, 4\}$ be the third party who can punish the investor $i \in \{1, 2\}$. If third party $k(i)$ reduces $P_k^i$ points from investor $i$'s payoffs, $k$ must pay $1/4^{\text{th}}$ of the reduced points from their 20 points. In other words, four times the number of points $k(i)$ paid will be deducted from $i$'s payoffs. Therefore, the maximum points by which $k(i)$ can reduce $i$'s payoffs is 80. However, $k(i)$ cannot reduce $i$'s payoffs by higher than $\pi_i^1$. Of the 20 points, those that $k(i)$ does not use to reduce $i$'s payoffs will be $k(i)$'s payoffs. Thus, the payoffs $\pi_i^I$ of investors and those $\pi_k^T$ of third parties from both the stages are given respectively by:

$$\pi_i^I = \max\left\{\pi_i^1 - P_{k(i)}^i, 0\right\}, \qquad for\ i = 1,2, \tag{2}$$

$$\pi_k^T = 20 - \frac{1}{4}P_k^{i(k)}, \qquad for\ k = 3,4, \tag{3}$$

where $i(k)$ is the inverse function of $k(i)$.

## 2.2. Roles of each subject

In our experiments, all subjects were assigned the roles of both an investor and a third party: they made decisions in those roles in each of the two decision stages of the underlying game as follows. At the beginning of Stage 1, all the subjects were divided into groups consisting of two subjects. Each subject received 100 points and had to choose investment points for the joint account that they shared with another subject in the same group. In Stage 2, all the subjects received an additional 20 points, regardless of the results of Stage 1; besides, they could use these 20 points to reduce the payoffs of another subject in a different group in Stage 1 as a third party. Thus, the experimental rewards for each subject $i$ were the sum of the payoffs as an investor and a third party.

For each subject $i$, the subject in $i$'s group (i.e., $i$'s pair), who $i$ could punish, and who could punish $i$ were determined in advance according to the identification numbers (IDs). Furthermore, the following four groups were different: the group to which the subject who $i$ could punish belonged, the group to which the subject who could punish $i$ belonged, the group to which the subject who could punish $i$'s pair belonged, and the group to which the subject who $i$'s pair could punish belonged.

At the beginning of Stage 2, for each subject $i$, their payoffs in Stage 1 were determined. However, each subject was not informed of this, so the results of Stage 1 had as little influence on their punitive behavior in Stage 2 as possible.

## 2.3. Strategy methods

To detect the conditional contributions of the subjects as an investor and the conditional punishment of the subjects as a third party, we adopted the strategy methods in both Stages 1 and 2. In Stage 1, we adopted the following strategy method similar to Fischbacher et al. (2001) : Each subject had to make two types of

decisions, one was regarding "unconditional investment" and the other was regarding "conditional investments." The subjects had to make both types of decisions without knowing the decisions of the others.

When it came to the "unconditional investment," subject $i$ had to decide $g_i \in \{0, 20, 40, 60, 80, 100\}$ points they wanted to invest in the joint account shared with their pair (say subject $j$). When it came to the "conditional investment," subject $i$ had to decide conditional investment $g_i(g_j)$ for each possible investment $g_j \in \{0, 20, 40, 60, 80, 100\}$ that subject $j$ could choose as their "unconditional investment." After the end of Stage 1, the computer randomly selected one of the two investors in the group. The "unconditional investment" was adopted as the actual investment of one investor selected by the computer, and the "conditional investment" was adopted as the actual investment of the other investor. The probability that subject $i$'s realized investment points will be determined by the unconditional investment decision is 50%, and that these points will be determined by the conditional investment decision is 50%. For example, if subject $i$ was selected by a computer, the total points in their joint account were $g_i + g_j(g_i)$.

To detect the conditional punishment of the subjects, we also adopted the following strategy method in Stage 2. At the start of Stage 2, the points which the subject who $i$ could punish (say, subject $l$) and their pair (say, subject $m$) had invested in their joint account have been confirmed. However, without knowing the realized investment amounts of $l$ and $m$, subject $i$ had to decide how much to reduce $l$'s payoffs for a combination of the possible investment amounts of $l$ and $m$. While the possible combinations of investment amounts of $l$ and $m$ in Stage 1 had 36 ($=6 \times 6$) patterns in total, we gave each subject only the following 12 patterns to mitigate their workload. That is, subject $i$ had to decide the points $P_i^l(g_l, g_m)$ to reduce $\pi_l^1$ ($l$'s payoffs in Stage 1) for the combinations of $l$'s all possible investments $g_l \in \{0, 20, 40, 60, 80, 100\}$ and $m$'s two types of investments $g_m \in \{a, b\}$ : $a$ is $m$'s realized investment points in Stage 1 and $b$ is the points that the computer had randomly chosen from five possible points other than $a$. For example, if $m$'s realized investment points were 40, then the computer selected one from $\{0, 20, 60, 80, 100\}$ at random. Of the 12 patterns, $P_i^l(g_l, g_m)$ of the points that subject $i$ selected for $(g_l, g_m)$ that had been realized in Stage 1 were deducted from $l$'s payoffs in Stage 1, and subject $i$ had to pay $1/4^{\text{th}}$ of $P_i^l(g_l, g_m)$ from the additional 20 points.

## 2.4. Treatments and hypothesis

Our experimental design consisted of two treatments: baseline (without moral messages) and nudge (with a moral message). Both treatments were essentially the same, except that the following message was written in the instructions and displayed on the computer screen in Stage 1 of the nudge treatment.

You should consider not only your profits but also the profits of your partner and try to invest a significant number of points into the joint account.

This study explores how this simple moral message influences the social preferences of the subjects regarding the payoff distributions of two investors and their willingness to punish the investors as a third party.

Several experimental studies have shown that certain types of nudge messages increase average investments in public goods games. For example, in Barron and Nurminen's (2020) experiment, the subjects were told that if everyone contributed to public goods, it would be beneficial for the group in their nudge treatment. Furthermore, the amount of investment above a certain level was labeled "good" written in green color and below that level was labeled "bad" written in red color in their instructions. Their results showed that the average investment in the nudge treatment was significantly higher than in the baseline treatment

without a nudge-based message.

These experimental results, however, do not imply that a subject's preferences for payoff distributions are affected by a nudge message because subjects' willingness to invest may increase with the average investment of other group members, as shown by Fischbacher et al. (2001). That is, if the subjects' expected amount for others' investments is increased due to the moral message, the subjects' investment will also increase without changing their preferences. Therefore, we focus on conditional investment $g_i(g_j)$ rather than unconditional investment to investigate the effects of a simple moral message on subjects' social preferences. Hence, the following hypothesis is proposed:

**Hypothesis 1: On average, the conditional investments in a nudge treatment are the same as in the baseline treatment.**

Our main concern is: whether a simple moral message affects subjects' willingness to punish as third parties. Note that if realized investments are increased by the simple moral message, third parties' punishments imposed on the investors in the nudge treatment may be less than those in the baseline treatment. Therefore, we must compare the conditional punishment between a nudge treatment and the baseline treatment. Thus, we test the following hypothesis.

**Hypothesis 2: On average, the conditional punishment pattern in a nudge treatment is the same as in the baseline treatment.**

## 2.5. Procedure

Our experiments were conducted at the Dokkyo University, Japan, between December 2021 and January 2022. We recruited subjects by displaying posters and distributing fliers. A total of 78 undergraduate students from several university departments participated in our study; these students had not participated in any prior public goods experiments. Furthermore, each subject could only participate in one treatment. The total number of subjects for the baseline treatment was 36 (12 per session $\times$ 3 sessions) and for the nudge treatment was 42 (12 per session $\times$ 2 sessions $+$ 18 $\times$ 1 session). For all the sessions, we used the z-tree software package developed by Fischbacher (2007).

Each session was conducted in a computer room. The subjects were seated in front of a computer terminal *at random*. Each desk was surrounded by partitions and had an envelope containing all the experimental materials, including instructions, a recording sheet, practice problems, and an ID card. IDs were also randomly assigned to each desk in advance so that the subjects could not predict which ID they will be assigned. During the experiment, the subjects could use the calculator placed on their desk any time.

To avoid potential experimenter effects, assistants other than the researchers acted as the instructors. The instructors read the instructions aloud so that the rules of the underlying game, the roles of each subject in Stages 1 and 2, and how their roles were determined were common knowledge for all the subjects. Before the subjects were introduced to the details of the decision tasks with strategy methods, they were instructed to solve practice problems to verify their understanding of the experimental design. The instructors began to read aloud the rest of the instructions that explained the details of the actual decision tasks with the strategy methods in Stages 1 and 2 after all the subjects provided correct answers to practice problems. Before the experiment began, we gave the subjects about 15 minutes to read the instructions on their own. Furthermore, to ensure thoughtful decisions, we did not impose a time limit on each decision task.

The subjects made decisions only once for each task. That is, there were no repetitions. For the subjects, the reward was a fixed participation fee of 1,000 yen plus points earned during the experiment.[1] One point was converted into 12 yen (0.25 points = 3 yen), and the reward was paid in cash to each subject after the session. The session lasted for approximately 90 minutes, and the subjects earned an average of 2,540 yen each.

## 3. Results

### 3.1. Overview

Before presenting the results of testing the two hypotheses, we present the descriptive statistics in Table 1. Although there were 78 subjects, we did not use the data from one subject because that subject gave up answering the practice problems. We find a large gap in the mean of the unconditional investment between the baseline and nudge treatments. The gap was significantly confirmed by the Wilcoxon rank-sum test ($p < 0.01$), implying that the moral message encouraged the subjects to invest more by influencing their social preferences. It was also observed that the same test above shows a significant difference in the means of the conditional punishment between the baseline and nudge treatments ($p < 0.01$). Therefore, the moral message may facilitate behavioral change such that the subjects punish more.

However, it is difficult to accept these results. This is because we employed the strategy method in the experiment, that is, the same subject appeared six or twelve times in the dataset. This data structure must be considered in the empirical analysis. In addition, individual characteristics, such as gender and faculty, may also affect subjects' conditional investment and punishment. With these considerations in mind, econometric analysis is conducted in the next section.

The total profit is shown in the last row of Table 1. The unit is not yen, but point. At first glance, the mean of the profit in the baseline treatment was lower than that in the nudge treatment. However, the Wilcoxon rank-sum test did not show a statistically significant difference ($p = 0.101$).

Male, economics, and living alone are dummies that take the value of 1 if the subject is male, a student in the faculty of economics, the subject's monthly income is in that range of its variable name, and the subject lives alone. The variable "Volunteer" refers to the number of times volunteers participated to date.[2] The "Favorite for math" is a five-point scale variable, where 1 and 5 indicate that math is the most and worst favorite subject, respectively.

1 1 US dollar was equal to approximately 115 yen in January 2022.
2 Types of volunteers are not specified.

Table 1：Descriptive statistics

| | ALL | | Baseline treatment | | Nudge treatment | |
|---|---|---|---|---|---|---|
| Variables | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Conditional investment[1)] | 28.05 | 32.76 | 19.90 | 28.59 | 34.84 | 34.49 |
| Conditional punishment[2)] | 7.61 | 14.58 | 6.42 | 13.71 | 8.59 | 15.21 |
| Male | 0.47 | 0.50 | 0.46 | 0.51 | 0.48 | 0.51 |
| Age | 20.30 | 1.28 | 20.57 | 1.14 | 20.07 | 1.35 |
| Economics | 0.44 | 0.50 | 0.51 | 0.51 | 0.38 | 0.49 |
| Income: none | 0.22 | 0.42 | 0.26 | 0.44 | 0.19 | 0.40 |
| Income: between 10 thousand and 20 thousand yen | 0.01 | 0.11 | 0.00 | 0.00 | 0.02 | 0.15 |
| Income: between 1 and 10 thousand yen | 0.04 | 0.19 | 0.06 | 0.24 | 0.02 | 0.15 |
| Living alone | 0.32 | 0.47 | 0.37 | 0.49 | 0.29 | 0.46 |
| Volunteer | 1.99 | 1.92 | 1.60 | 1.80 | 2.31 | 1.97 |
| Favorite for math | 2.36 | 1.16 | 2.49 | 1.17 | 2.26 | 1.15 |
| Total profit | 128.34 | 24.53 | 123.71 | 24.09 | 132.20 | 24.51 |

Note: 1) The number of observations is 462, 210, and 252 in ALL, the baseline treatment, and the nudge treatment, respectively, because these variables are derived from the strategy method.

Note: 2) The number of observations is 924, 420, and 504 in ALL, the baseline treatment, and the nudge treatment,, respectively, because each subject chooses punishment 12 times according to their third-party investments. For the variables below male, the number of observations is 77, 35, and 42 for ALL, the baseline treatment, and the nudge treatment, respectively. These numbers are the same as the number of subjects.

### 3.2. Tests for Hypothesis 1

To examine Hypothesis 1, we regress a subject's conditional investment on a dummy variable for the nudge treatment and other control variables. The number of observations was 462 (77 subjects × six types of pairs' investment exogenously given) because the data used in the regression were obtained from the strategy method. The standard errors clustered in subjects were applied to account for individual effects. The estimation results are presented in Table 2.

It is significantly found that the subjects in the nudge treatment invest, on average, more than those in the baseline treatment. This result is consistent with the descriptive statistics presented in Table 1. The difference between the two treatments may not be constant, as in Models (1) and (2). To consider this potential, Model (3) contains the interaction term with the nudge treatment and the pair's unconditional investment as explanatory variables. The coefficient of the interaction term is significant at the 5% level, implying that the effect of the moral message increases with the pair's investment instead of remaining constant.

The coefficients of the pair's investment in the strategy method are significantly positive at the 1% level in all the models, suggesting that the more the pair invested, the more they invested themselves. The magnitudes of the coefficients are less than 0.5, indicating that the subjects invested less than half of the pair's investment. Regarding the results of the other control variables, male (female) is likely to invest less (more).

Table 2 : Results for the conditional investment

| | (1) | (2) | (3) |
|---|---|---|---|
| Nudge | 14.94*** | 13.86*** | 1.540 |
| | (4.980) | (4.195) | (4.184) |
| Nudge × Pair's investment | | | 0.246** |
| | | | (0.102) |
| Pair's investment | 0.452*** | 0.452*** | 0.318*** |
| | (0.0522) | (0.0527) | (0.0740) |
| Male | | -11.41** | -11.41** |
| | | (4.804) | (4.810) |
| Age | | 0.398 | 0.398 |
| | | (1.804) | (1.806) |
| Economics | | -6.967 | -6.967 |
| | | (4.767) | (4.772) |
| Income: none | | -6.630 | -6.630 |
| | | (4.728) | (4.733) |
| Income: between 10 thousand and 20 thousand yen | | -44.97*** | -44.97*** |
| | | (4.329) | (4.334) |
| Income: between 1 and 10 thousand yen | | 14.07* | 14.07* |
| | | (7.615) | (7.623) |
| Living alone | | -2.008 | -2.008 |
| | | (4.875) | (4.880) |
| Volunteer | | 1.291 | 1.291 |
| | | (1.306) | (1.308) |
| Favorite for math | | -2.565 | -2.565 |
| | | (2.100) | (2.102) |
| Constant | -2.693 | 3.865 | 10.58 |
| | (2.995) | (36.71) | (36.85) |
| F-value | 86.67*** | 23.91*** | 23.45*** |
| Adusted R-squared | 0.27 | 0.35 | 0.37 |

Note: The number of observations is 462. Standard errors are clustered by subjects. ***, **, and * correspond to the one, five, and ten percent level of significance, respectively.

### 3.3. Tests for Hypothesis 2

We hypothesized that the moral message would influence a subject's punitive behavior. To examine this hypothesis, we regressed conditional punishment on the nudge dummy. The number of observations was 924 (77 subjects × 12 types of subjects' investments who are third party). Standard errors were clustered by subjects. We employed a Tobit regression in addition to the ordinary least squares (OLS) method used in the previous subsection 3.2. This is because the number of observations with zero punishment was 567 (approximately 61% of the total). In addition to the control variables such as gender, we added behavior dummies for punitive behavior, which were created by a cluster analysis of the 12 types of punishment. The 77 subjects were classified into three types: less punishment (60), moderate punishment (10), and more punishment (7). The less-punishment type was used as the criterion, and dummy variables for the other two types were employed in the analysis.

The estimation results are displayed in Table 3, where the OLS and Tobit models are used in Models (1) to (3) and (4) to (6), respectively. In all estimations, no significance was confirmed for the coefficients of the nudge message, whereas we observed statistically significant coefficients of the interaction term with differences in investment in the other designated pair, as in Models (3) and (6). These results imply that the impact of the moral message is not fixed; however, as the difference increases, the moral message becomes

more influential in the direction of punishment. The more unequal the behavior of the two subjects in the third party, the more a subject punishes them, even if they accept the cost of reducing their own profits. The nudge will reinforce this trend. As for the other control variables, some of them were found to be significant, but they were inconsistent.

Table 3 : Results for third party's punishment

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | | OLS | | | Tobit | |
| Nudge | -0.0426 | -0.221 | -0.00193 | 0.451 | -0.476 | -0.636 |
| | (1.018) | (1.081) | (1.079) | (3.229) | (3.305) | (3.308) |
| Nudge × Difference of investment in the other pair | | | 0.0452** | | | 0.0914** |
| | | | (0.0182) | | | (0.0410) |
| Difference of investment in the other pair | 0.0359*** | 0.0375*** | 0.0124 | 0.0912*** | 0.0974*** | 0.0442* |
| | (0.00983) | (0.0102) | (0.00769) | (0.0220) | (0.0231) | (0.0250) |
| Male | | -0.340 | -0.379 | | -2.584 | -2.513 |
| | | (0.932) | (0.936) | | (3.264) | (3.237) |
| Age | | -0.377 | -0.372 | | -1.369 | -1.340 |
| | | (0.354) | (0.360) | | (1.248) | (1.227) |
| Economics | | 1.264 | 1.226 | | 4.530 | 4.593 |
| | | (1.061) | (1.068) | | (3.383) | (3.379) |
| Income: none | | -1.550 | -1.689* | | -5.130 | -5.257 |
| | | (0.961) | (0.972) | | (3.572) | (3.612) |
| Income: between 10 thousand and 20 thousand yen | | -3.797 | -4.474 | | -11.93** | -12.82** |
| | | (3.646) | (3.734) | | (5.182) | (5.234) |
| Income: between 1 and 10 thousand yen | | -1.151 | -0.947 | | -0.909 | -1.001 |
| | | (1.187) | (1.171) | | (6.113) | (6.081) |
| Living alone | | 0.126 | 0.218 | | 1.130 | 1.320 |
| | | (1.161) | (1.163) | | (3.537) | (3.506) |
| Volunteer | | -0.0314 | -0.0232 | | 0.706 | 0.717 |
| | | (0.220) | (0.222) | | (0.758) | (0.752) |
| Favorite for math | | -0.647 | -0.618 | | -2.641* | -2.627* |
| | | (0.422) | (0.422) | | (1.357) | (1.353) |
| Cluster: moderate punishment | 17.63*** | 17.32*** | 17.36*** | 30.64*** | 29.99*** | 30.08*** |
| | (1.538) | (1.713) | (1.676) | (3.338) | (3.855) | (3.871) |
| Cluster: more punishment | 37.68*** | 38.60*** | 38.52*** | 50.97*** | 54.86*** | 54.57*** |
| | (3.242) | (3.563) | (3.622) | (4.268) | (5.297) | (5.156) |
| Constant | 2.032*** | 11.33 | 10.93 | -11.90*** | 20.59 | 19.78 |
| | (0.705) | (7.482) | (7.588) | (3.426) | (25.08) | (24.60) |
| F-value | 459.5*** | 143.81*** | 136.81*** | 41.34*** | 41.36*** | 43.66*** |
| Adusted/Pseudo R-squared | 0.67 | 0.67 | 0.67 | 0.16 | 0.18 | 0.18 |

Note: Number of observations is 924. Standard errors are clustered by subjects. ***, **, and * correspond to the one, five, and ten percent level of significance, respectively.

## 4. Concluding remarks

This study experimentally explores how a simple moral message that reminds subjects of altruism influences people's social preferences and willingness to punish as third parties in a simple public goods game with third parties' punishment. We found that the simple moral message used in the study enhanced subjects'

reciprocity and their willingness to punish as third parties those who made unfair investments. However, the total profits were not significantly different between the nudge and baseline treatments.

The lack of an effect of the simple moral message on total payoffs may be due to simultaneous positive and negative factors. The positive factor was cooperation and the negative factor was altruistic punishment. While the simple moral message promoted altruistic and cooperative behavior, it accelerated altruistic punishment in situations. As these effects cancel each other out, the nudge eventually had no impact on social welfare.

Our experimental results suggest that moral messages are more effective in situations where deviant behavior is easily discovered by a third party. For example, it is easy for a third party to observe whether people are wearing masks and the self-restraint of businesses at restaurants, but it is difficult for a third party to determine whether a person has been vaccinated unless the person confesses. A systematic analysis of the relationship between the observability of deviant behavior and the effectiveness of simple moral messages is an important topic for future research.

## References


Allcott, H. 2011. Social norms and energy conservation. *Journal of Public Economics*, 95(9-10), 1082-1095.

Barron, K., and Nurminen, T. 2020. Nudging cooperation in public goods provision. *Journal of Behavioral and Experimental Economics*, 88, 101542.

Bicchieri, C., and Dimant, E. 2022. Nudging with care: the risks and benefits of social information. *Public Choice*, 191(3), 443-464.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), 1015-1026.

Ferraro, P. J., and Price, M. K. 2013. Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *The Review of Economics and Statistics*, 95(1), 64-73.

Fehr, E., and Fischbacher, U. 2004. Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2), 63-87.

Fischbacher, U., Gächter, S., and Fehr, E. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters*, 71(3), 397-404.

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

Hallsworth, M., Chadborn, T., Sallis, A., Sanders, M., Berry, D., Greaves, F., et al. 2016. Provision of social norm feedback to high prescribers of antibiotics in general practice: A pragmatic national randomised controlled trial. *The Lancet*, 387(10029), 1743-1752.

Sasaki, S., Kurokawa, H., Ohtake, F. 2021. Effective but fragile? Responses to repeated nudge-based messages for preventing the spread of COVID-19 infection. *The Japanese Economic Review*, 72, 371-408.