Cooperation among behaviorally heterogeneous players in a social dilemma with stay or leave decisions

Xiaochuan Huang, Takehito Masuda[*,†,‡], and Tatsuyoshi Saijo[§,**]

October 2017

Abstract

We experimentally test a two-stage mechanism called the *stay-leave mechanism* to achieve cooperation in *n*-player prisoner dilemma situations. Each player who cooperates can revise his choice when players' choices are not unanimous. All players cooperate by eliminating dominated strategies in each stage under our mechanism. The result holds when each player is either strategic or conditionally cooperative. The cooperation rate in the mechanism experiment averaged 86.6% across 15 periods and 96% after period 5. Elimination of the dominated strategies, conditional cooperation and pessimistic defection coexist in such a way that partially explains the observed upward trend in the cooperation rate.

JEL Classification Codes: C72; C72; D74; H41; P43

Keywords: social dilemma; experiment; mechanism; behavioral heterogeneity

[*] Corresponding Author. E-mail: takehitomasuda@gmail.com

[†] Institute of Social and Economic Research, Osaka University, Mihogaoka 6-1, Ibaraki, Osaka 567-0047 Japan

[‡] Economic Science Laboratory, Eller College of Management, University of Arizona, 1130 East Helen Street, McClelland Hall 401, Tucson, AZ 85721-0108, USA

[§] Research Institute for Humanity and Nature, 457-4 Motoyama, Kamigamo, Kita-ku, Kyoto, 603-8047 Japan

[**] Research Center for Future Design, Kochi University of Technology, 2-2 Eikokuji, Kochi 780-8515, Japan

1. Introduction

Since the conventional model of voluntary public goods provision was presented by Bergstrom, Blume, and Varian (1986), many studies have attempted to understand human subjects' motivation to free ride (Andreoni, 1988; Isaac, Schmidtze, & Walker, 1989; Ledyard, 1995; Zelmer, 2003). Additional studies have attempted to design the rules of games (mechanisms) that provide players with incentives while causing individual selfish choices to result in efficient outcomes (Chen, 2008). Nonetheless, designed mechanisms often deviate from efficient outcomes with human subjects partly due to individuals' heterogeneous motives beyond assumed selfish motives (see, e.g., Charness, Fréchette, & Qin et al., 2007; Levati & Neugebauer, 2004).[1] Thus, the fundamental challenge that naturally arises is to design mechanisms that align with heterogeneous behavioral rules.

Saijo, Masuda, Okano, and Yamakawa (2016) shed light on this issue by developing an approval mechanism for prisoners' dilemmas. After having played a prisoner's dilemma, players approved or rejected the other's choice of cooperation or defection. If both players approved the other's choice, the outcome was simply the result of the chosen strategies in the prisoner's dilemma. However, if either rejected the other's choice, the outcome was the result of mutual defection in the prisoner's dilemma. The authors observed a consistent cooperation rate of 93.2% with backward elimination of weakly dominated strategies (BEWDS). Additionally, classified individual choices revealed that the driving force behind subjects' behavior is heterogeneous including reciprocity and inequity aversion other than BEWDS.[2] Given these results, Masuda,

[1] Charness et al. (2007) reported, in their Varian's (1994) compensation mechanism experiment with asymmetric prisoner's dilemmas, that the preference for the equity of the final payoffs hinders reaching the subgame perfect Nash equilibrium. Levati and Neugebauer (2004) reported, in their English auction to determine contributions to public goods, that selfish players who dropped out at the beginning of the auction triggered conditionally cooperative player dropouts.

[2] See an earlier version Saijo, Okano, and Yamakawa (2016). The authors also showed experimentally

Okano, and Saijo (2014) designed a public goods mechanism based on the same behavioral rule and experimentally verified that it works well. However, neither Saijo et al. (2016) nor Masuda et al. (2014) explored the *n*-player case. A common logic behind their mechanism is that when a refund occurs privately upon individual request, the cooperator (the player who chose a larger contribution) has incentive to do so, resulting in a reversion to low aggregate contributions, which is perceived as a threat by the defector (the player who chose a smaller contribution).[3]

Using the same logic in the *n*-player prisoner's dilemma environment, this paper proposes a two-stage mechanism called the *stay-leave mechanism* (SLM), which proceeds as follows. In the first stage, each player chooses Cooperation (*C)* or Defection (*D*). If all choose *C* or all choose *D*, the game ends, and the corresponding first-stage choices are implemented. Otherwise, after observing the other players' choices, only players who chose *C* in the first stage can proceed to the second stage where they choose to *Stay* or *Leave*. If a player chooses to *Stay*, that player contributes the endowment. If the player chooses to *Leave*, that player makes no contribution. No *D* player proceeds to the second stage and, therefore, the *D* player contributes nothing.

Our main prediction is that the SLM attains cooperation as a unique BEWDS equilibrium for any marginal per capita return that arises in the *n*-player prisoner's dilemma (in other words, the SLM *implements* cooperation in BEWDS, Proposition 1).[4] Then, the implementation result is

that BEWDS provided a better prediction across different mechanisms compared to the Nash equilibrium and the subgame perfect Nash equilibrium.

[3] In contrast, imposing an exogenous threshold on total contributions (Croson and Marks, 2000) showed that provision for an ex-post unanimity voting stage (Banks, Plott, & Porter, 1988; Dannenberg, 2012; Fischer & Nicklisch, 2007) to approve contributions is likely to create a coordination problem. Brams and Kilgour (2009) used one-stage unanimous voting between mutual cooperation and mutual defection, which will fail to maximize welfare if either player is so altruistic. On the other hand, approval mechanisms, where every outcome of the dilemma can occur, can deal with such situations.

[4] Choosing "approval" may change the outcome of the game in an institutional design approach by Saijo et al. (2016) and Masuda et al. (2014). However, this is not true in a strand of "social approval" literature that empirically examines the effect of ex-post appraisal of others' contributions on the outcomes in public goods games (see, e.g., Masclet, Noussair, Tucker, & Villeval, 2003).

extended to players who are heterogeneous in behavior (Proposition 2). In this case, we assume players are composed of BEWDS players and conditional cooperators based on the stylized fact that conditional cooperators prevail frequently in dilemma experiments (see Arifovic and Ledyard, 2011; Chaudhuri, 2011 for an overview). Conditional cooperators "cooperate if there is sufficient chance that their opponent will do likewise" (Andreoni & Samuelson, 2006).

We conduct experiments with groups of three and random matching. We find that introducing the SLM significantly increases average final cooperation rates compared to the *n*-player prisoner's dilemma only. In the SLM sessions using the direct method, cooperation rates averaged 86.6% when we combined the data across all 15 periods while they averaged 96% after period 5. In contrast, the *n*-player prisoner's dilemma-only sessions yielded an average cooperation rate of only 18.5%. However, our original theory cannot explain the observed upward trends in cooperation rates, particularly at the beginning of the SLM.

We find evidence of behavioral heterogeneity associated with this trend. Our analysis of individual choices in the SLM sessions with the strategy method and belief elicitation revealed that 47.1% and 14.9% of subjects, respectively, were deemed to follow BEWDS and conditional cooperation. Additionally, first-stage defection with a pessimistic wait-and-see stance triggered others to switch to defection by choosing *Leave*, and explains eventual no group contribution in period 1.

Our contributions are as follows. First, our primary experimental result of high cooperation rates in the SLM contributes to a strand of experimental literature that aims to design effective mechanisms using classic equilibrium concepts, particularly for similar refund mechanisms (Coats, Gronberg, & Grosskopf., 2009; Croson & Marks, 2000) for discrete public goods and Isaac et al., 1989; Gerber & Wichardt, 2009 for the linear public goods). Second, we contribute to the emerging literature that induces cooperation in a social dilemma (and that is not limited to an institutional design approach) admitting the spectrum of conditional cooperation (Andreroni & Samuelson, 2006; Levati & Neugebauer, 2004; Steiger & Zultan, 2014) since we provide a simple model of a behaviorally mixed group.[5] This approach seems more fruitful because by considering our third contribution, we observe stable prevalence of BEWDS subjects and conditional cooperators in line with the literature that mainly explores subjects' motives in a voluntary contribution experiment (Arifovic and Ledyard, 2014; Chaudhuri, 2011; Fischbacher, Gachter, & Fehr, 2001; Steiger and Zultan, 2014).

The remainder of this paper is organized as follows. Section 2 provides our model, and Section 3 offers the main prediction. Section 4 explores behavioral heterogeneity. Section 5 describes the experimental design, Section 6 discusses the experimental results, and Section 7 concludes.

2. The Stay-leave mechanism

In this section, we present some preliminaries and state our main theoretical result. To show the intuitiveness of our solution, we begin with a public good provision with two players. Each player $i = 1, 2$ is endowed with \$10 and must decide whether to contribute \$10 to the public good (denoted by C) or consume \$10 privately (denoted by D). The sum of the contribution is multiplied by $\alpha = 0.7,$ and non-rivalness ensures that the benefit of the public good passes to every player. The game has a prisoner's dilemma game structure. Both players' contributions maximize the sum of the payoffs yielding (14, 14). However, individual interests conflict with

[5] Although we consider the SLM a variant of the IF*n* treatment of Gerber, Neitzel, & Wicardt (2014), one novelty of this paper is to examine how behavioral heterogeneity affects the outcomes in an institution and provide evidence that behavioral heterogeneity prevails in the institution.

those of the collective. Because a player’s contribution renders the player worse off by $3(=10-7=17-14)$ regardless of what the other player does, no contribution occurs at the dominant strategy equilibrium (*D*, *D*) yielding (10, 10).

We consider a simple mechanism so that the unique equilibrium outcome is cooperative (14, 14), that is, the SLM. Under the SLM, a cooperator has the chance to revise their choice when players’ choices are not unanimous (see Figure 1).

Figure 1 around here.

In the first stage, players simultaneously and privately choose *C* or *D*. If both choose *C*, the game ends. Moreover, the outcome or players’ payoff vector is $(14,14)$. If player 1 chooses *C* but player 2 chooses *D* (i.e., *CD*),[6] only player 1 proceeds to the second stage and decides whether to *Stay* in cooperation or *Leave* to defection. If player 1 chooses *Stay* at *CD*, the outcome is the players’ choice in the first stage, $(7,17)$. In contrast, if player 1 chooses *Leave* at *CD*, the outcome when both defect is (10, 10). According to the symmetric argument, in subgame *DC*, if player 2 chooses *Stay*, the outcome is (17, 7). However, if player 2 chooses *Leave*, the outcome is (10, 10). Finally, if both choose *D*, the game ends, and both receive 10.

We mention a few comparisons of the SLM with mechanisms in the related literature. First, a noteworthy feature of the SLM is that it does not use private punishment technologies, which are sometimes assumed in the literature (Fehr & Gächter, 2000; Kamei, Putterman, & Tyran, 2015; Masclet et al. 2003; Noussair & Tan 2011). However, since personal punishment is

[6] Hereafter, subgames are indexed by *n* letters of *C* or *D* unless the index is confusing. Moreover, if the players’ identity does not matter, we put *C’s* index first. For example, we write *CCCD* when $n = 4$.

typically prohibited in modern societies or legal systems, we view this as a strength of the SLM. On the other hand, under the SLM, when a player chooses to *Leave*, the money is simply returned to the player.

Second, since the SLM refunds privately upon request, the SLM potentially improves welfare compared to a simple refund conditional on total contributions less than an exogenous threshold Coats et al., 2009; Croson and Marks, 2000; Isaac et al., 1989). To see this, suppose there are three players, and the first-stage choices are *CDD*. Then, the public good will not be provided at all under a simple refund with an exogenous threshold of two *C*s while the public good may be provided under the SLM. Later, we return to the welfare perspective after investigating the data exhibiting behavioral heterogeneity.[7]

3. Theoretical prediction

In this section, we show that all players cooperate in the unique equilibrium of BEWDS.

We solve the game presented in Figure 1 using BEWDS. Consider subgame *CD*. Player 1 compares 7 and 10 and, then, chooses *L*. The same holds for player 2 for subgame *DC*. By incorporating the subgame outcomes, we can construct the reduced normal form game shown in Table 1. Then, the pair payoff player 1 obtained by choosing *C* is [14, 10] while that obtained by choosing *D* is [10, 10]. Since [14, 10] weakly dominates [10, 10], player 1 chooses *C*. The same is true for player 2. Thus, the unique outcome is (14, 14).

Table 1 around here.

[7] Other related literature includes an experiment on mechanisms enacted endogenously by players (Kosfeld, Okada, & Riedl, 2009; Kube, Schaube, Schildberg-Horisch, & Khachatryan, 2015) while we ignore the participation problem for the SLM.

Following Bergstrom (2002), we formulate the social dilemma (SD) as an *n*-player public good provision with binary choices for $n \geq 2.$ Each player $i = 1, 2, ..., n$ endowed with $w > 0$ units of the private good chooses *C* or *D*. If $k \geq 0$ players choose *C*, all *n* players receive the benefit of the public good, $\alpha kw,$ where $1/n < \alpha < 1.$ Additionally, each *D* player also receives the benefit from private consumption. Then, the total payoff is maximized when all players choose *C* yielding $(\alpha nw, ..., \alpha nw)$, and this is called unanimous cooperation hereafter. However, for any $k \leq n-1$, that is, regardless of what the other players choose, a player would choose *D* to increase the payoff by $\{\alpha kw + w\} - \alpha(k+1)w = (1-\alpha)w.$ That is, the dominant strategy is *D*. Hence, no public good is provided in an SD-only setting.

The extension of the SLM to the multi-player case is simple. In the first stage, players simultaneously and privately choose *C* or *D*. If all players choose *C* or all choose *D*, the game ends, and the corresponding first-stage choices are implemented. Otherwise, all *C* players proceed to the second stage and simultaneously and privately decide to *Stay* or *Leave*. If the *C* player chooses *Stay*, the player finally contributes *w*. If the *C* player chooses *Leave*, the player contributes nothing. No *D* player proceeds to the second stage and, thus, a *D* player contributes nothing. Now, we obtain the following result.

Proposition 1. *Assume* $n \geq 2.$ *The SLM implements cooperation in BEWDS.*

Proof. See Appendix.

4. Behavioral heterogeneity and misbeliefs about type explain the delay in cooperation

It is a stylized fact that the subjects participating in social dilemma tend to “cooperate if there is sufficient chance that their opponent will do likewise” (Andreoni & Samuelson, 2006). Motivated by this fact, we extend the model so that players are a mixture of BEWDS and a conditional cooperator who chooses *C* in the first stage, chooses *Stay* believing the other *C* player will choose *Stay* in *CCD*, and chooses *Leave* in *CDD*.

Example. Consider $n = 3, \alpha = 0.7.$ Suppose that player 1 uses BEWDS while players 2 and 3 are conditionally cooperative, which is known by player 1. Table 2 shows the subjective reduced normal form game for player 1. We omit player 3’s *C* since it is obvious. The light gray area indicates that player 1 need not consider the area because player 1 knows both the other players are conditionally cooperative. Player 1 also expects players 2 and 3 to choose *Stay*, mutually observing their Cs. Given this, player 1’s payoff by choosing *D* is 24, the highest payoff that player 1 can earn. Hence, *D* is not weakly dominated by *C* for player 1. Since player 1 chooses *D* while player 2 and 3 chooses *C* then *Stay*, each obtains 24, 14, and 14.

-------------------------

Table 2 around here.

-------------------------

The example suggests that we cannot incentivize all players to cooperate if conditionally cooperative players are a large fraction of the players. We formalize this intuition in what follows. For *n*-players, a player is *conditionally cooperative* if that player satisfies the following conditions:

a) the player chooses *C* in the first stage;

b) the player chooses *Leave* in subgame *CD…DD*; and

c) the player chooses *Stay* unless in *CD…DD* believing that some *C* player will choose *Stay*.

d) the player believes that at least one of the other players will choose *C*

Now, assume that $n \geq 2$ players consisting of $c \in \{1, 2, ..., n-1\}$ conditionally cooperative players and $(n - c)$ BEWDS players. For simplicity, we assume that each BEWDS player $i \leq n - c$ knows *c*. To ensure weak dominance, we also assume

$$\alpha \in A \equiv (1/n, 1) \setminus \{1/(n\text{-}1), 1/(n\text{-}2), ..., 1/2\}. \qquad (1)$$

Let *G*(*n*,*c*) denote the set of heterogeneous groups *g* such that there are $c \in \{1, 2, ..., n-1\}$ conditional cooperators and (*n*-c) BEWDS players. Henceforth, given n and *c*, we say the SLM *almost implements cooperation in heterogenous groups on* $G(n, c)$ if, for any $\alpha \in A$ and any $g \in G(n, c)$, all players in group *g* cooperate with probability one in the unique predicted outcome of the game specified by n, α, c and the SLM. Then, we have the following result.

Proposition 2. *(i) Assume n=2. The SLM almost implements cooperation when a group consists of one BEWDS player and one conditional cooperator.*

(ii) Assume $n \geq 3$*. The SLM almost implements cooperation in heterogenous groups on* $G(n, c)$ *if and only if* $c \leq \lfloor n - 1/\alpha \rfloor$ *where* $\lfloor x \rfloor$ *is the largest integer that does not exceed x.*

Proof. See Appendix.

5. Experimental design

We conducted experiments on the SLM sessions and SD sessions as a control at Osaka

University in October 2012, January and March 2013, and March 2016.[8] Readers can refer to the online supplementary information to see all experimental materials.

Basic design across treatments

We set $n = 3$ and $\alpha = 0.7$ and used a random matching protocol. In every period, each subject was given 1,000 experimental currency units (ECUs). That is, if all three group members choose *D*, they each received 1,000 ECUs. In each session, subjects played the game for 15 periods. No individual participated in more than one session. Subjects were recruited from Osaka University through campus-wide advertisements. We used z-Tree software (Fischbacher, 2007).

Subjects were seated at computer terminals, separated from each other with partitions. Communication was prohibited among subjects. Each subject received written instructions and record sheets (see supplementary materials). An experimenter read the instructions aloud, and subjects were then given five minutes to ask questions. In each period, subjects were anonymously divided into groups of three. We informed the subjects of the random matching. After finishing all 15 periods, subjects were asked to complete a questionnaire, immediately after which they were privately paid in cash. Subjects were paid an amount proportional to the sum of ECUs that they had earned over the 15 periods. Table 3 summarizes the experimental design.

[8] We also conducted sessions as an extension of Saijo et al.'s (2016) unanimous voting type for the approval mechanism but omitted the data because the sessions were beyond the scope of this paper. If all approve the others' choices, the outcome is simply the result of the chosen strategies in the dilemma. However, if either disapproves, the outcome is the result of all defection in the dilemma. As we mentioned in a previous footnote, such a unanimous voting mechanism fails to achieve cooperation in BEWDS, theoretically and experimentally. See also the footnote in Section 6 and Huang, Masuda, Okano and Saijo. (2015) for the approval mechanism data.

Table 3 around here.

SLM sessions using the direct method (SLM-direct)

The SLM-direct treatment continued as follows. In the first stage (called the choice stage in the experiment) of each period, by observing the payoff matrix, each subject was asked to select either *C* or *D*, which were presented using the neutral labels *B* and *A*, respectively, in the experiment and to mark their choices along with the reason for their choice in the record sheet. Once all subjects finished their tasks, they were instructed to click the *OK* button.

Then, subjects observed the first-stage choices of their group and whether they would proceed to the second stage (called the new choice stage in the experiment). If the first-stage choices were *CCC* or *DDD*, group members proceeded to the result screen, which is explained later. Otherwise, each *C* player proceeded to the second stage. In the second stage, by observing the payoff matrix, *C* players were asked to select either *Stay* or *Leave* ("stay with *B*" or "change to *A*" in the experiment) and input their choice into the computer. They were then asked to write down their choices along with the reason in the record sheet. In contrast, *D* players could not proceed to the second stage, and they were asked to wait for the others.

Once all subjects who proceeded to the second stage had finished the procedure and clicked the *OK* button, everyone proceeded to the result screen. The result screen included the first-stage choices, the *C* players' second-stage choices, and each group member's earnings in the period. After all subjects wrote down their earnings and clicked the *Next* button, the following period began. No information on the choices of the other groups was provided to subjects. There was no practice period.

Prior to these tasks, subjects answered non-incentivized pre-play questionnaires at the beginning of each period regarding their choices and their opinion as to what their group member choices would be in the first-stage and second-stage subgames. Although there are six second-stage subgames in total, because of the symmetry of the other two players, it was sufficient to ask about four subgames, *CCD, CDD, DCC,* and *DCD*, where the first character indicates the responder's own choice in the first stage. After participants completed the questionnaire, they proceeded to the first stage.[9] Finally, the SD treatment did not include a second stage.

SLM sessions using the strategy method (SLM-strategy)

To check the robustness of the results obtained under SLM-direct, we performed the following additional treatments in March 2016. The first-stage was the same as in SLM-direct. Then, each subject who chose C in the first stage in the SLM-strategy responded to questions concerning their choice plan in both subgames *CDD* and *CCD* before learning what other group members chose in the first stage.

Moreover, every subject answered, regardless of their first-stage choice, non-incentivized *mid-play questionnaires* set to elicit their belief on others' choices in the first-stage and relevant second-stage subgames. Relevant subgames are *CDD* and *CCD* for each *C* player while relevant subgames are *DCD* and *DCC* for each *D* player.[10] Once other group members' planned actions were revealed, the relevant choices were realized. We also set a practice period before the actual experiment to help subjects understand the rules, which seemed less intuitive compared with SLM-direct, and to minimize any effect on the payment periods.

[9] For the list of questions, see the supplementary material. Before the second stage, subjects also answered questionnaires concerning their hypothetically choices and what they think *C* players would choose in the subgame the group actually reached. We did not find notable results for this questionnaire.

[10] For the list of questions, see the supplementary material.

6. Experimental results

Average cooperation rates

Figure 2 shows the time path of the average cooperation rate over the 15 periods sorted by treatment. We use the cooperation rates after the second stage (henceforth, the cooperation rates for simplicity) in the SLM-direct and SLM-strategy.

Result 1. *The average cooperation rate in SLM-direct averaged 86.6% across 15 periods with an upward trend.*

The average cooperation rate in the SLM-direct sessions (the line with the circle symbols) was 44.4% in the first period. Across all 15 periods and three sessions, subjects in the SLM cooperated, on average, 86.6% of the time. Out of the 315 observed group outcomes in the SLM (7 groups × 15 periods × 3 sessions), all three players cooperated in 268 observations. Focusing on the time after period 5, the average cooperation rate increased to 96%. In fact, the Spearman's rank correlation test indicates convergence to the cooperative outcome, showing that the upward time trend in the average cooperation rate under the SLM was statistically significant $(p = 0.040)$.

Figure 2 around here.

The SD sessions (the line with the square symbols) replicated the observed pattern of previous SD experimental studies. In the first period, subjects cooperated 19% of the time, and this rate gradually decreased to 4.8% in the last period. The overall average cooperation rate of the SD was 18.6%. Overall, just three of the 210 groups achieved a cooperative outcome. The downward trend in the average cooperation rate was statistically significant (Spearman's rank correlation test; $p < 0.001$).

For the SLM-strategy sessions (the line with the diamond symbols), in period 1, the second-stage cooperation rate averaged 41.4%, similar to SLM-direct. Across 15 periods, the second-stage cooperation rate averaged 65%. Spearman's rank correlation test supports the increasing time trend of the cooperative outcome $(p < 0.001)$.[11]

Result 2. *The SLM significantly increased the average cooperation rate compared to the SD. However, the strategy method had a significantly negative impact on the average cooperation rate.*

We perform the Mann–Whitney test, following Andreoni and Miller's (1993) analysis of prisoner's dilemma experiments. That is, we first calculate each subject's average cooperation rate across 15 periods. Then, we regard each subject's average as a one-unit observation and run the tests. We find that the SLM significantly increases the average cooperation rate regardless of the elicitation method (*p*-values for SLM-direct vs. SD < 0.001; *p*-values for SLM-strategy vs. SD < 0.001) There is significant difference between two elicitation methods (*p*-values for SLM-direct vs. SLM- strategy < 0.001).

Behavioral type classification using the strategy method

[11] Huang et al. (2015) reported that the cooperation rates in the AM averaged 57.7%.

For a better understanding of what happened in the SLM sessions, particularly in the beginning periods, we classified each subject into one of four behavioral types. We focused on period 1 data to assign a unique behavioral type for each subject and to avoid the effects of interactions among subjects. We mainly present two arguments: the coexistence of BEWDS and conditional cooperators, and a low group cooperation rate in response to pessimistic defection.

Table 4 shows the prevalence of subjects' behavioral types and the corresponding criteria based on information elicited with incentive in the SLM-strategy sessions.

-----------------------

Table 4 around here.

-----------------------

The first category is BEWDS, which requires a subject (i) to choose *C* in the first stage, (ii) to choose *Leave* whenever the subject can, and (iii) to expect the other *C*-player to choose *Leave* in *CCD*. The second category is the conditional cooperator, in a group of three, who answered (i), (ii)' to choose *Leave* if they were a unique *C*-player, and (iii)' to choose *Stay* expecting the other C-player to choose *Stay* in *CCD.*

The third and fourth classifications are for the *D*-player. Optimistic defection captures the motivation to exploit, demanding (i)' to choose *D* expecting the other two players to choose *CC* and (ii) to expect mutual *Stay* from the other two *C*-players. Finally, pessimistic defection captures the motivation to wait-and-see. This type (i)'' chooses *D* believing that at least one of the group members chooses *D* and (iii) expects the C-player to choose *Leave*. Note that our classification criteria cannot be sufficient conditions for each behavior due to a lack of information on the unachievable second stages. We obtain the following result.

Result 3. *Subjects in the SLM-strategy sessions exhibited behavioral heterogeneity. (i) In particular, two-thirds of the data in period 1 are explained by BEWDS subjects and conditionally cooperative subjects. (ii) Observed full defection in the group was mainly triggered by pessimistic defection. (iii) Moreover, the behavioral type distribution was not significantly different between the elicitation methods.*

Figure 3 shows the prevalence for each behavioral type in the first period of the SLM-strategy and SLM-direct. Each colored area in the stacked bar chart represents subjects classified as *BEWDS*, conditionally cooperative, optimistic defection, and pessimistic defection. Interestingly, subjects were heterogeneous in behavior. In the SLM-strategy, BEWDS explains 47.1% of the data while conditional cooperation explains 14.9%. Remarkably, optimistic defectors (5.8%) are only approximately half of pessimistic defectors belief that other(s) will choose *D* (10.3%).

Figure 3 around here.

These two types of *D*-players explain most failures to achieve cooperation in period 1 because the existence of either type of *D*-player in a group triggers contributors' refund in stage 2 and, hence, collapses all contributions during the period (i.e., *CCD* or *CDD,* then, all *C*-players chose *Leave*). Such cases account for 66.7% (=8/12) and 90% (=9/10) of the zero-group contribution at the end of period 1 for the SLM-strategy and SLM-direct. The observed share of behavior is not significantly different between the two elicitation methods (*p* = 0.801, chi-squared test).

7. Concluding remarks

We introduced the SLM for *n*-player prisoner dilemma games to achieve cooperation in BEWDS and mixed populations consisting of BEWDS and conditional cooperation. Under the SLM, each cooperator has the chance to revise their choice when players' choices are not unanimous. In our SLM experiment, we observed convergence with the cooperative outcome after period 5 with an average cooperation rate of 96%. Moreover, our analysis of elicited individual choices in the strategy method revealed BEWDS and that conditionally cooperative players coexist.

The implications of our experiment shed light on the importance of incorporating behavioral heterogeneity into institutional design to achieve cooperation in line with Andreoni and Varian (1999), Charness et al. (2007), and Levati and Neugebauer (2004). One fruitful way to do this is to consider mechanisms for a continuum of conditionally cooperative players proposed in Andreoni and Samuelson (2006) where both selfish and altruistic players are extreme cases.

The current study has limitations and future directions. The first limitation is that we do not compare the SLM directly with other mechanisms in the literature such as the compensation mechanism. Our observation contrasts with the results of previous experimental studies, such as Varian's (1994) compensation mechanism experiment for two-player prisoner's dilemma games (Andreoni & Varian, 1999; Charness et al., 2007), where the cooperation rate reached approximately 70% after dozens of repetitions.

Second, as we mentioned in Section 2, the SLM can improve welfare compared to a simple refund with an exogenous threshold (e.g., Isaac et al., 1989) under behavioral heterogeneity. Since the SLM refunds only when players request it, the mechanism can respect the intention of conditional cooperators and altruistic players to cooperate even if there are some free-riders. On the other hand, the threshold mechanism cannot respect these intentions. To elaborate on this point, however, we need sophisticated modeling of conditional cooperators similar to Andreoni and Samuelson (2006) and Steiger and Zultan (2014), so that we can examine heterogeneous groups in a wide range of mechanisms.

Acknowledgements

The authors are very grateful to Yoshitaka Okano for his contributions to previous versions of this paper. We thank Takako Greve, Chiaki Hara, Tadashi Sekiguchi, and Kazumi Shimizu. Keiko Takaoka also provided outstanding research assistance. We also thank the participants of the game theory seminar at Kyoto University, Asia Economic Institutes Workshop, the Meeting of the Japanese Society for Mathematical Biology, and GAMES 2016. This research was supported by JSPS KAKENHI Grant Number 24243028, K-CONNEX, and “Experimental Social Sciences: Toward Experimentally-based New Social Sciences for the 21st Century,” a project under the aegis of the Grant-in-Aid for Scientific Research on Priority Areas of the Ministry of Education, Science, and Culture of Japan.

Appendix

Proof of Proposition 1. Let $n \geq 2$ and $\alpha \in (1/n, 1)$. Assume that all players use BEWDS. Consider first any second-stage subgame of the SLM after $n-1$ or less players chose *C* in the first stage. Pick any player who choses *C*. Then, by construction of the SLM, the second-stage mover chooses *Leave* because it gains $(1-\alpha)w$ rather than *Stay*. Then, this player receives w in this subgame. Next, consider the reduced normal form game. By the above argument, the reduced normal form game is such that each player will receive αnw if all *n* players choose *C* in the first stage, and each player will receive w otherwise. Moreover, $\alpha nw > w$ by $\alpha \in (1/n,1)$. Therefore, *C* weakly dominates *D* in the first stage.

Proof of Proposition 2. Without loss of generality, we assume any $i \le n - c$ is a BEWDS player. (i) We consider only two cases: *CC* and *DC*. If player 1 chooses *D*, player 2, the conditional cooperator, chooses *Leave* in the second stage. Hence, player 1 obtains *w* by choosing *D.* Hence, player 1 will choose *C* and obtain $2\alpha w$.

(ii) $(\Rightarrow)$. Take any $n \ge 3$, $\alpha \in A$. It suffices to consider BEWDS player 1. Consider the subgame where all players except for player 1 choose *C*. If Player 1 chooses *C*, the player obtains a payoff of $n\alpha w$. If player 1 chooses *D*, since there are two or more *C* among (*n*-1) other players, every conditionally cooperative player chooses *Stay*. Thus, player 1 obtains a payoff of $w + c\alpha w$. Thus, *C* weakly dominates *D* only if $n\alpha w \ge w + c\alpha w$. Since *c* is an integer and by assumption (1),

$$c \le \lfloor n - 1/\alpha \rfloor \text{ and } n\alpha > 1 + c\alpha \qquad (2)$$

$(\Leftarrow)$ Suppose next that $c \le \lfloor n - 1/\alpha \rfloor$. By the above argument, *C* yields a higher payoff when all other players choose *C* (*). There are two cases where player 1 can or cannot affect others' choices. Note that by $\lfloor n - 1/\alpha \rfloor \le n - 1/\alpha < n - 1$.

Case 1 $c = 1$. Consider the subgame where only a unique conditional cooperator chooses *C* while all other players choose *D*. In this subgame, player 1 can induce the conditional cooperator's *Stay* by choosing *C* while he cannot by choosing *D*. Hence, player 1 can increase his payoff by α. Together with (*), this implies that *C* weakly dominates *D.*

Case 2 $2 \le c \le n - 2$. Since there are two or more conditional cooperators, all of them chooses *C* and then *Stay* regardless of the choice of BEWDS players in the first stage. Hence, player 1 is indifferent between *C* then *Leave* and *D* for any subgame where the number of C players is between *c* and *n*-2. Together with (*), this implies that *C* weakly dominates *D*.

References

Andreoni, J. (1988). Why free ride? Strategies and learning in public goods experiments. Journal of public Economics, 37(3), 291-304.

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. The Economic Journal, 103(418), 570–585.

Andreoni, J., & Samuelson, L. (2006). Building rational cooperation. Journal of Economic Theory, 127, 117–154.

Andreoni, J., & Varian, H. (1999). Preplay contracting in the prisoners' dilemma. Proceedings of the National Academy of Sciences of the United States of America, 96(19), 10933–10938.

Arifovic, J., & Ledyard, J. (2011). A behavioral model for mechanism design: Individual evolutionary learning. Journal of Economic Behavior & Organization, 78, 374–395.

Banks, J. S., Plott, C.R., & Porter, D.P. (1988). An experimental analysis of unanimity in public goods provision mechanisms. Review of Economic Studies, 55(2): 301-322.

Bergstrom, T. C. (2002). Evolution of social behavior: Individual and group selection. The Journal of Economic Perspectives, 16(2), 67-88.

Bergstrom, T. C., Blume, L., &. Varia, H. (1986). On the private provision of public goods. Journal of Public Economics, 29(1), 25–49.

Brams SJ, Kilgour DM (2009) How democracy resolves conflict in difficult games. In: Levin S (ed) Games, Groups, and the Global Good. Springer, Berlin, pp. 229-241.

Charness, G., Fréchette, G.R., & Qin, C.-Z. (2007). Endogenous transfers in the Prisoner's Dilemma game: An experimental test of cooperation and coordination. Games and Economic Behavior, 60(2), 287–306.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. Experimental Economics, 14(1), 47–83.

Chen, Y., (2008). Incentive-compatible mechanisms for pure public goods: A survey of experimental literature. In Plott, C.R., & Smith, V.L. (Eds.), The Handbook of Experimental Economics Results, Elsevier, 625–643.

Coats, J. C., Gronberg, T. J., & Grosskopf, B. (2009). Simultaneous versus sequential public good provision and the role of refunds—an experimental study. Journal of Public Economics, 93(1), 326-335.

Croson, R.T.A., & Marks, M.B. (2000). Step returns in threshold public goods: A meta- and experimental analysis. Experimental Economics, 2, 239-259.

Dannenberg, A. (2012). Coalition formation and voting in public goods games. Strategic Behavior and the Environment, 2(1), 83-105.

Fehr, E., & Gächter, S. (2000) Cooperation and punishment in public goods experiments. American Economic Review, 90(4): 980-994.

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

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

Fischer, S., & Nicklisch, A. (2007). Ex interim voting: An experimental study of referendums for public-good provision. Journal of Institutional and Theoretical Economics, 163(1), 56-74.

Gerber, A., Neitzel J., & Wichardt., P.C. (2013). Minimum participation rules for the provision of public goods. European Economic Review, 64, 209-222.

Gerber, A., & Wichardt, P. C. (2009). Providing public goods in the absence of strong institutions. Journal of Public Economics, 93(3), 429-439.

Huang X, Masuda T, Okano Y, Saijo T (2015). Cooperation among behaviorally heterogeneous players in social dilemma with stay or leave decisions. KUT-SDE working papers 2014-7, Kochi University of Technology.

Isaac, R. M., Schmidtz, D., & Walker, J. M. (1989). The assurance problem in a laboratory 173 market. Public choice, 62(3), 217-236.

Kamei, K., Putterman, L., & Tyran, J. R. (2015). State or nature? Endogenous formal versus informal institutions in the voluntary provision of public goods. Experiential Economics, 18 (1), 38–65.

Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games, American Economic Review, 99, 1335-1355.

Kube, S., Schaube, S., Schildberg-Hörisch, H., & Khachatryan, E. (2015). Institution formation and cooperation with heterogeneous agents. European Economic Review, 78, 248-268.

Levati, M.V., & Neugebauer, T. (2004). An application of the English clock market mechanism to public goods games. Experimental Economics, 7, 153–169.

Ledyard, J.O. (1995). Public goods: a survey of experimental research, in: Kagel J., Roth, A. (Eds.), The Handbook of Experimental Economics, Princeton University Press, Princeton, pp. 111-194.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. The American Economic Review, 93(1), 366-380.

Masuda, T., Okano, Y., & Saijo, T. (2014). The minimum approval mechanism implements the efficient public good allocation theoretically and experimentally. Games and Economic Behavior 83, 73–85.

Noussair, C. N., & Tan, F. (2011). Voting on punishment systems within a heterogeneous group. Journal of Public Economic Theory, 13(5), 661-693.

Saijo, T., Masuda, T., Okano, & Y., Yamakawa, T. (2016). Approval mechanism to solve prisoner’s dilemma: Comparison with Varian’s compensation mechanism. KUT-SDE working papers 2016-15, Kochi University of Technology.

Saijo, T., Masuda, T., Okano, & Y., Yamakawa, T. (2016). The approval mechanism experiment: A solution to prisoners dilemma. KUT-SDE working papers 2015-12, Kochi University of Technology.

Steiger, E.-M., & Zultan, R. (2014). See no evil: Information chains and reciprocity. Journal of Public Economics, 109, 1-12

Varian, H. (1994). A solution to the problem of externalities when agents are well-informed. American Economic Review, 84(5), 1278–1293.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. Experimental Economics, 6(3), 299-310.

Tables and Figures

Table 1. Reduced normal form game under the SLM

		Player 2	
		C	*D*
Player 1	*C*	14,14	10,10
	D	10,10	10,10

Table 2. Payoff table for example 1

		Player 2	
		(conditional cooperator)	
		C	*D*
Player 1	*C*	21,21,21	17,17,7
(BEWDS)	*D*	24,14,14	10,10,10

Table 3. Summary of the experimental design.

Treatment	Behavior/belief elicitation	Payment scheme	Number of sessions (subjects)
SLM-direct	Direct method/pre-play questionnaire[a)]	Total	3 (63)
SLM-strategy	Strategy method/mid-play questionnaire[a)]	Total / single period[b)]	6 (87)
SD	-	Total	2 (42)

Notes. a) For the list of questions, see Section 1 of the supplementary material. All pre-/mid-play questionnaires on others' actions are non-incentivized. b) Since there is no significant difference in the average first-stage cooperation rates between total and single payment sessions, we merged the data for both payment schemes.

Table 4. Criteria for behavioral classification

		Behavioral classification			
Response		BEWDS	Conditional cooperation	Optimistic defection	Pessimistic defection
Own stage 1 choice		*C*	*C*	*D*	*D*
Own choice plan	$\underline{C}DD$[a]	*Leave*	*Leave*	N.A.	N.A.
in stage 2 after	$\underline{C}CD$	*Leave*	*Stay*	N.A.	N.A.
	$C\underline{C}D$	*Leave*	*Stay*	N.A.	N.A.
Guess on choice(s)	$D\underline{C}D$	N.A.	N.A.	*Leave*	*Leave*
in stage 2 after	$D\underline{CC}$	N.A.	N.A.	*Stay,* *Stay*	*Leave,* *Leave*
Guess on stage 1 choices		N.R.	N.R.	*CC*	*DD* or *CD*

Notes. N.A. = Not Available, N.R. = Not Restricted. [a] The underline indicates the subject who was asked to describe their actions.

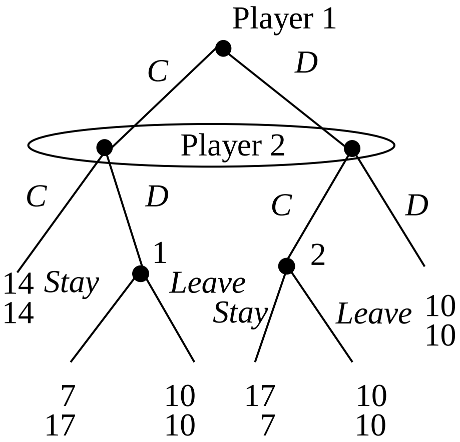

Figure 1. The SLM

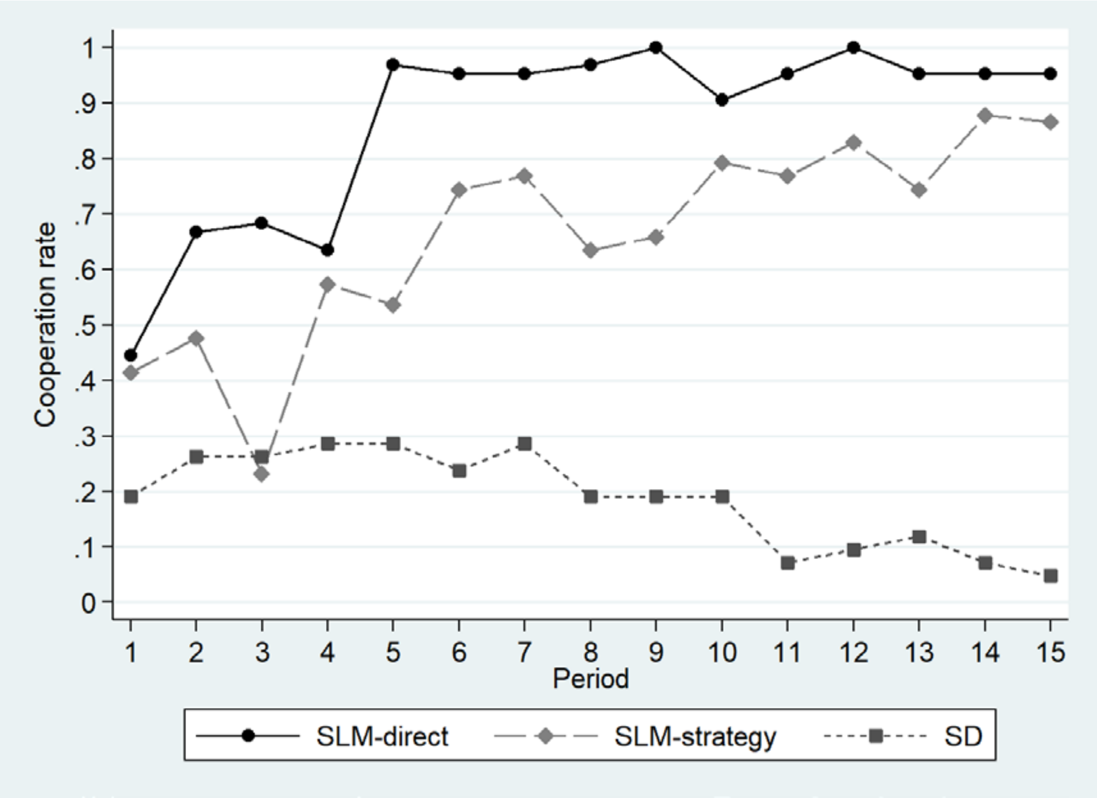

Figure 2. Average cooperation rate after the second stage by period and sorted by mechanism

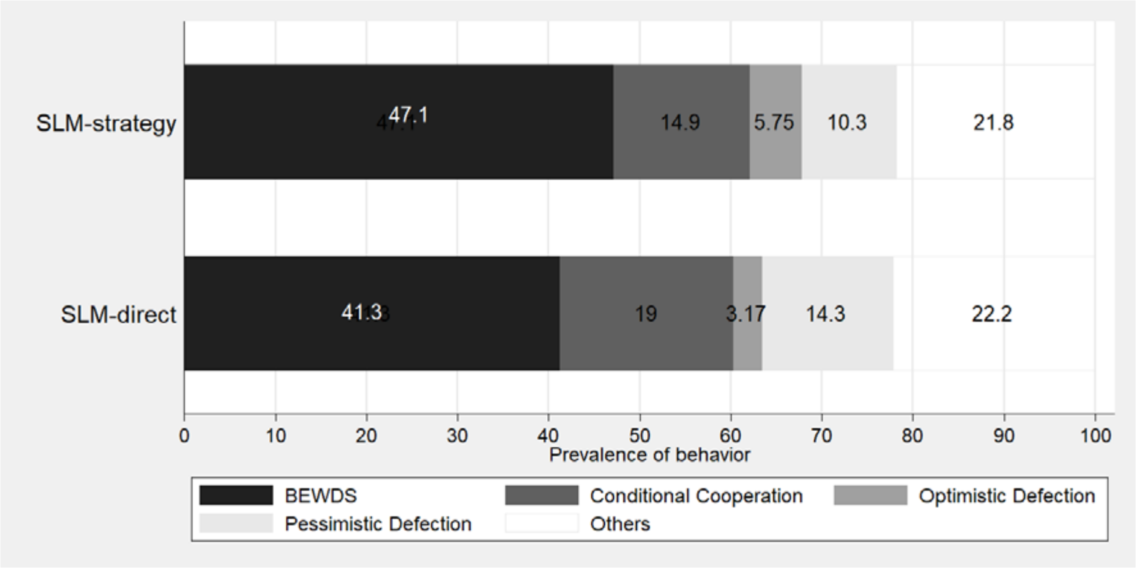

Figure 3. Percentages of behavioral type in period 1