KOCHI UNIVERSITY OF TECHNOLOGY

Social Design Engineering Series *SDES-2014-7*

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

Xiaochuan Huang
DT Captical Management Co., LTD.

Takehito Masuda
Research Center for Social Design Engineering, Kochi University of Technology
Japan Society for the Promotion of Science

Yoshitaka Okano
Kochi University of Technology
Research Center for Social Design Engineering, Kochi University of Technology

Tatsuyoshi Saijo
Kochi University of Technology
Research Center for Social Design Engineering, Kochi University of Technology
Center for environmental Innovation Design for Sustainability, Osaka University
Institute of Economic Research, Hitotsubashi University

26th February, 2015

School of Economics and Management
Research Center for Social Design Engineering
Kochi University of Technology

KUT-SDE working papers are preliminary research documents published by the School of Economics and Management jointly with the Research Center for Social Design Engineering at Kochi University of Technology. To facilitate prompt distribution, they have not been formally reviewed and edited. They are circulated in order to stimulate discussion and critical comment and may be revised. The views and interpretations expressed in these papers are those of the author(s). It is expected that most working papers will be published in some other form.

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

Abstract

Given the substantial evidence of behavioral heterogeneity in social dilemma experiments, in this study we consider how to achieve cooperation in *n*-player prisoner's dilemma situations where each player has one behavioral type, either selfish or conditionally cooperative. We introduce a two-stage game form called the stay-leave mechanism, where each cooperator has the chance to revise his choice when players' choices are not unanimous. For homogeneous behavioral type cases, theory predicts that the unique outcome is cooperative under the stay-leave mechanism, regardless of the benefit derived from cooperation. For heterogeneous behavioral type cases, we show that the benefits of cooperation shrink such that the unique outcome is cooperative under the stay-leave mechanism if the number of conditionally cooperative players increases. The average cooperation rate in the stay-leave mechanism experiment is 86.6% across 15 periods, which increases to 96.0% after period 5. We also provide evidence that selfish and conditionally cooperative types coexist throughout the sessions, in the proportion that the equilibrium outcome is cooperative. Our results corroborate the importance of incorporating behavioral heterogeneity into institutional design.

JEL Classification Codes: C72; C92; D74; H41; P43

Keywords: social dilemma; experiment; conditional cooperator; behavioral heterogeneity

1. Introduction

A long-standing question in economics is how to foster cooperation when conflicts exist between individuals and collectives. The canonical formulation of such a social dilemma is the voluntary provision of a public good. A large body of knowledge addresses this question by changing the rules of games using a money transfer, dynamic structure, and so on to incentivize players to achieve a cooperative outcome. Players are typically assumed to be selfish and have homogeneous behavioral rules, but heterogeneous payoffs.[1]

Nevertheless, growing experimental evidence has suggested the limitations of applying such an approach. First, assumed rationality may be demanding even though it approximates the data after a sufficient duration of learning. Andreoni and Varian (1999) and Charness et al. (2007) show that even in simple two-player asymmetric prisoner's dilemma games, subjects need repetition to learn the cooperative subgame perfect Nash equilibrium under the compensation mechanism proposed by Varian (1994).

Moreover, subjects in laboratory experiments have heterogeneous behavioral rules even when payoffs are symmetric. Andreoni and Samuelson (2006), Chaudhuri (2011), and Arifovic and Ledyard (2011) mention a developing consensus in (typically symmetric) social dilemma experiments that a significant proportion of subjects are *conditionally cooperative*, that is, individuals "cooperating if there is sufficient chance that their opponent will do likewise" (Andreoni and Samuelson, 2006). Gächter (2007) and Croson and Shang (2008) summarize the field evidence on such conditional cooperators.[2]

As such evidence accumulates, theories to enhance cooperation under the existence of conditional cooperators are emerging. Specifically, the new research line of institutional design regards behavioral rules, instead of monetary payoffs, as each player's characteristic feature, herein called his or her *type*.[3] Levati and Neugebauer (2004), for example, propose applying an English auction to linear public good provision so that conditionally cooperative players commit to contribute to the public good unit by unit to achieve the efficient outcome. However, the auction in their experiments does not work, since selfish types' drop out at the low contribution level trigger other conditionally cooperators' drop out. Andreoni and Samuelson (2006) consider how to weight the payoffs in two-period prisoner's dilemma games to increase cooperation among types of conditional cooperators, but do not aim to achieve the unique efficient equilibrium. Likewise, Grimm and Mengel (2009), who consider the endogenous sorting of selfish and conditionally cooperative types,

[1] For the details of public good mechanism design and various implementation approaches, see Groves and Ledyard (1977), Ledyard and Palfrey (1999), and Chen (2008). Chen and Plott (1996) and Healy (2006) observe that subjects require feedback on other players' choices to adjust to the Nash equilibrium.

[2] See also Brandts and Scram (2001), Cason and Gangadharan (2014), Chaudhuri et al. (2006), Ledyard (1995), and Kurzban and Houser (2005).

[3] We use the word "type" in accordance with Gächter (2007).

allow subjects to have hundreds of repeated choices in two prisoner's dilemma games with different gains for defectors. However, situations such as charitable fundraising are subject to constraints, as reported in Croson and Shang (2008), are not so frequently repeated as in Andreoni and Varian (1999) and Grimm and Mengel (2009) and operated by designers without coercive power, who rather rely on the voluntary cooperation of players. Hence, inducing cooperation among behaviorally heterogeneous players under such constraints is an interesting and significant research topic.

In this spirit, the present study explores the possibility of cooperation, within a few repetitions, in symmetric n-player prisoner's dilemma situations when each player has one of two behavioral types: selfish or conditionally cooperative. In our baseline social dilemma setting, each player endowed with one unit of a private good decides to contribute (C after cooperation) or not (D after defection) the endowment to the public good. One additional C generates the benefit of $\alpha \in (1/n, 1)$ to every player, under which all-D is the inefficient dominant strategy equilibrium.[4]

To achieve the cooperative outcome, we introduce a two-stage game form called the *stay-leave mechanism* (SLM). The SLM proceeds as follows. In the first stage, each player chooses C or D. After observing the other players' choices, only players who chose C in the first stage can proceed to the second stage, where they choose *Stay* or *Leave*. If a player chooses *Stay*, he contributes the endowment. If the player chooses *Leave*, he makes no contribution. Roughly speaking, under the SLM, the fundraiser grants each contributor a chance to receive a refund.

We assume that selfish types eliminate weakly dominated strategies in each subgame. We call such a behavioral rule backward elimination of weakly dominated strategies (BEWDS), a concept that originated from Kalai (1981) and that has since been experimentally reexamined by Saijo et al. (2015) and Masuda et al. (2014).[5] On the contrary, conditionally cooperative types choose C in the first stage; then, if the number of other players choosing C in the first stage is less than some threshold, the player chooses *Leave*, and *Stay* otherwise. We implicitly assume that conditionally cooperative types are myopic in the sense that they perceive another player's first-stage choice of C as a signal to choose

[4] Across disciplines, there have been decades of extensive research on this question using n-player prisoner's dilemma games as the simplest representation of voluntary contribution to a public good. See Schelling (1973), Hamburger (1973), Dawes (1980), Ledyard (1995), Ostrom (2000), and Nowak (2006) for examples.

[5] To our knowledge, Kalai (1981) is the first study to construct a mechanism by using BEWDS. Kalai (1981) adds negotiation steps prior to the n-player prisoner's dilemma game, with each step being at most n stages and where each player is asked whether to revise his choice (i.e., cooperation or defection) as long as he has not done so before. This process continues until every player has revised his choice or there are no players remaining that choose to revise. However, Kalai (1981) does not formally indicate that his negotiation mechanism leads to a unique cooperative outcome in BEWDS for some of the negotiation steps required for implementation.

Stay in the second stage.[6] For the sake of simplicity, we assume that type composition and the threshold of conditionally cooperative types are common knowledge.

We obtain two main theoretical results. First, in the homogeneous type case, when players are all selfish or all conditionally cooperative, the unique equilibrium outcome under the SLM is the cooperative one. The logic for the all-selfish case is simple. In any second stage, choosing *Leave* yields at least a payoff of 1, while *Stay* yields at least $\alpha < 1$, regardless of others' second-stage choices. The same holds for the other players. Thinking backwardly, a selfish player finds that every second stage yields a payoff of 1 except when all players choose *C*, which yields $\alpha n > 1$, and hence he chooses *C*. When all players are conditionally cooperative, on the contrary, the result is straightforward.

The second main result is for the heterogeneous type case for $n \geq 3$. Suppose m players are selfish and the remaining $n - m$ players are conditionally cooperative, where $m \in \{1, 2, ..., n-1\}$, and that these two types are distinguishable in the sense that in some decision node they choose differently. The condition such that the unique equilibrium outcome is cooperative is simply characterized as $\alpha \in [1/m, 1)$.

This condition has two interpretations. The first interpretation is that for three or more players with at least one conditional cooperator, it is not the case that the unique equilibrium outcome is cooperative for all social dilemma cases under the SLM because there exists $\alpha \in (1/n, 1/m)$. The second interpretation is that there is a tradeoff between the number of conditional players and the range of α within which the unique equilibrium outcome is cooperative under the SLM. Actually, type composition endogenously emerges and evolves through players' interactions and hence it is not common knowledge. In such an environment, selfish types' belief about others' types matters since the above second result suggests that selfish types' overestimation of the number of conditionally cooperative types decreases the cooperation rate. Hence, the above predictions are worth evaluating experimentally.

To this end, we conducted experiments to evaluate the degree to which the SLM enhances cooperation compared with the social dilemma setting and to elicit subjects' behavioral types. We chose the parameters $n = 3$ and $\alpha = 0.7$ since these parameters set theoretically can lead to both cooperation and non-cooperation cases. The cooperative outcome is achieved if two selfish types exist (or are expected to exist according to selfish types) in a group by $1/2 < 0.7$, whereas the cooperative outcome is not achieved if only one selfish type exists (or is expected to exist according to selfish types) in a group by

[6] For a more sophisticated formulation of conditionally cooperative players who care about others' past behavior or intention, see Rabin (1993) and Falk and Fischbacher (2006). Another approach is to assume that a player has the utility function representing the distributional concern (see Bolton and Ockenfels 2000).

$1 > 0.7$. We used a random matching protocol in the experiments.

In the SLM treatment, we find that the average cooperation rate is 86.6% when we combine the data across all 15 periods, while it is 96.0% after period 5. To explain why the SLM is suitable, we analyze group behavior and individual behavior. First, a closer look at group behavior supports the coexistence of both selfish and conditionally cooperative types. The selfish choice is *Leave* in every second stage. In all cases where only one subject in the group chose *C* in the first stage, the subject chose *Leave* in the second stage. For 37.9% of the groups where two subjects chose *C*, however, at least one subject chose *Stay* in the second stage.

Second, we identify the behavioral types of subjects in the SLM by classifying the pair of first-stage choices and the responses to the pre-period questionnaire to predict other group members' choices. We find evidence that selfish and conditionally cooperative types coexist throughout the session. Moreover, the estimated type composition is such that the cooperative outcome is achieved theoretically. In particular, in the first period of the SLM, 41.3% of subjects are revealed to be selfish, while 19.0% are conditionally cooperative. Over time, some shifts toward the selfish type occur. Although selfish-type players amount to 54.0% in periods 1–4 where the cooperation rate is less than 70%, this increases to 75.6% in periods 5–15 where the cooperation rate is 90% or more. On the contrary, conditional cooperators comprise 15.1% in periods 1–4, but only 10.0% in periods 5–15. At the same time, almost all selfish types expect another two players to also be selfish, roughly consistent with the empirical type share.

This study contributes to the literature on institutional design incorporating several behavioral rules. Levati and Neugebauer (2004) theoretically consider only behaviorally homogeneous cases of all-selfish or all-conditionally cooperative. Likewise, Masuda et al. (2014) consider several selfish types but assume that players have the same type for two-player linear public good provision. Masuda et al. (2014) find their mechanism works well in the experiment, and also show the evidence for behavioral heterogeneity by using the third-person classification of subjects' free-form answers. In this paper, on the other hand, we try to go one step further by considering any mixed population of selfish and conditionally cooperative types.

The remainder of this paper is organized as follows. Section 2 theoretically shows that cooperation between a population that comprises selfish and conditionally cooperative players can cooperate under the SLM. It is also shown that the tradeoff between the number of conditionally cooperative types and the range of value of cooperation within which the unique equilibrium outcome is cooperative. Section 3 describes the experimental design. Section 4 discusses the experimental results. Section 5 concludes.

2. The model

2.1. The SLM

In this section, we present some preliminaries and then 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 to contribute \$10 to the public good (denoted by *C*) or to 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' contribution maximizes the sum of the payoffs, yielding (14, 14). Nevertheless, individual interests conflict with those of the collective. Because a player's contribution makes 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 game form so that the unique equilibrium outcome is a cooperative one (14, 14), that is, the SLM. Under the SLM, a cooperator has the chance to revise his choice when players' choices are not unanimous (see Figure 1).

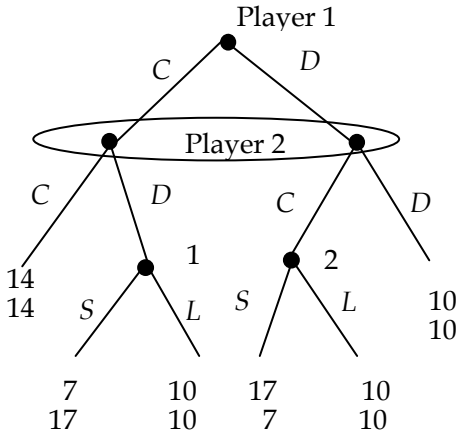

Figure 1. The SLM.

In the first stage, players simultaneously and privately choose *C* or *D*. If both choose *C*, the game ends; furthermore, the outcome or players' payoff vector is $(14,14)$. If player 1 chooses *C* but player 2 chooses *D* (i.e., *CD*),[7] only player 1 proceeds to the second stage and decides whether to stay (*S*) in cooperation or leave (*L*) to defection. If player 1 chooses *S* at

[7] 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 first. For example, we write *CCCD* when *n* = 4.

CD, the outcome is the players' choice in the first stage, $(7,17)$. On the contrary, if player 1 chooses *L* at *CD*, the outcome is that when both defect, (10, 10). According to the symmetric argument, in subgame *DC*, if player 2 chooses *S,* the outcome is (17, 7). However, if player 2 chooses *L,* the outcome is (10, 10). Finally, if both choose *D*, the game ends and both receive 10.

2.2. Behavioral homogeneity cases

In this subsection, we deal with cases when all players are selfish or all are conditionally cooperative under the SLM. We confirm that in both cases the unique outcome is cooperative. Let us begin with the all-selfish case. Saijo et al. (2015) develop a modified two-stage prisoner's dilemma game with the unique cooperative outcome under BEWDS. In their experiment with perfect stranger matching, the authors observe a cooperation rate of 93.2%. Saijo et al. (2015) also show experimentally that BEWDS provides a clear prediction compared with the Nash equilibrium and subgame perfect Nash equilibrium. Masuda et al. (2014) design a public good mechanism based on BEWDS and experimentally verify that it works well. Therefore, we use BEWDS as our basic behavioral principle for selfish types.

Next, we solve the game presented in Figure 1, assuming that all players use BEWDS. Consider subgame *CD*. Player 1 compares 7 and 10 and then chooses *L*. The same holds for player 2 at subgame *DC*. By incorporating subgame outcomes, we can thus construct the reduced normal form game shown in Table 1. Then, by comparing (14, 10) with (10, 10), each player chooses *C,* because the former weakly dominates the latter. Thus, the unique outcome is (14, 14) in BEWDS.[8]

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

Table 1. Reduced normal form game under the SLM.

The extension of the SLM to the many players case is simple. In the first stage, players simultaneously and privately choose *C* or *D*. If all choose *C* or all choose *D*, the game ends. Otherwise, all *C* players proceed to the second stage and simultaneously and privately decide *S* or *L*. If the *C* player chooses *S*, he finally contributes w. If the *C* player

[8] Iterative elimination of weakly dominated strategies yields the same result. First, since *CS* is dominated by *CL*, we eliminate *CS*. Second, since *D* is dominated by *CL*, then we eliminate *D*.

chooses *L*, he contributes nothing. Similarly, no *D* player proceeds to the second stage and thus contributes nothing. To understand that all choose *C* even if there are many players, see Table 2, which illustrates the subgames under the SLM when there are four players.

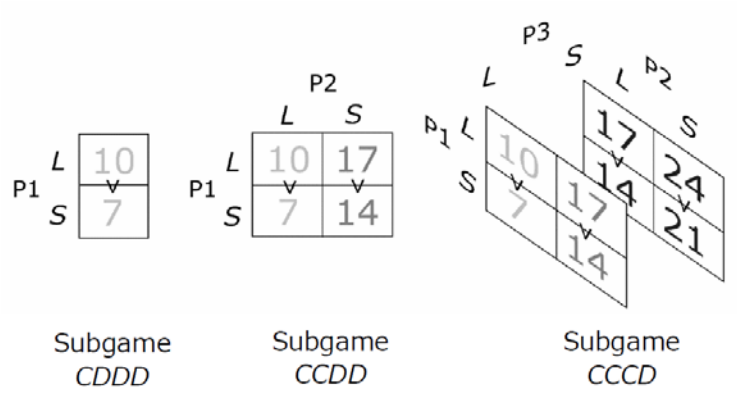

Table 2. Player 1's payoffs in subgames under the SLM.

Consider the first subgame *CDDD* on the left side of Table 2. As shown by the construction of the SLM, only player 1 proceeds to the second stage. After comparing 10 and 7, player 1 chooses *L*.

Consider the next subgame *CCDD* in the center of Table 2. The left column of subgame *CCDD* contains the same payoffs as those for subgame *CDDD*. Moreover, because *S* for player 2 means that player 2 contributes, each entry in the right column of the payoffs of subgame *CCDD* is larger than that in the left column by 7. Hence, choice *L* yields a higher payoff than *S* for player 1 even when player 2 chooses *S*. In sum, *L* dominates *S* for player 1. By using a symmetric argument, player 2 also chooses *L*.

Finally, consider subgame *CCCD* on the right side of Table 2. The left panel of subgame *CCCD* contains the same payoffs as those of subgame *CCDD*. Hence, for the left panel, *L* dominates *S*. Because *S* for player 3 means that player 3 contributes, each entry in the right panel of subgame *CCCD* is larger than that in the left panel by 7. The dominance relation is then preserved in the right panel of subgame *CCCD*. In sum, *L* dominates *S* in subgame *CCCD*. By using symmetric arguments, players 2 and 3 also choose *L*.

In summary, in every second-stage subgame, every *C* player chooses *L*. Then, the outcomes are (10, 10, 10, 10) unless the first-stage choices are *CCCC* where the outcome is (28, 28, 28, 28). We can therefore construct the reduced normal form game where *C* weakly dominates *D*.[9]

A similar argument holds in general. Before stating the first proposition, let us

[9] There are two subgame perfect Nash equilibrium outcomes: (28, 28, 28, 28) and (10, 10, 10, 10).

formulate SD as an n-player public good provision with binary choices. Each player $i \in I = \{1,2,...,n\}$, $n \geq 2$ endowed with one unit of the private good chooses C or D. If k players choose C, all n players receive the benefit of the public good, αk, where $\alpha \in (1/n, 1)$. In addition, each D player also receives the benefit from private consumption. Then, the total payoff is maximized when all players choose C, yielding $(\alpha n,...,\alpha n)$, called the cooperative outcome herein. However, regardless of what the other players choose, a player would choose D to increase his payoff by $\{\alpha(k-1)+1\} - \alpha k = 1 - \alpha$. That is, the dominant strategy is D. Hence, no public good is provided in a social dilemma only setting.

Denote i's choice by (s_i, t_i), where $s_i \in \{C, D\}$. In addition, if i is a C player, $t_i \in \{S, L\}$. Otherwise, set $t_i = -$ to indicate that i does not proceed to the second stage. Let x be the number of players choosing (C, S). Under the SLM, player i's payoff is given by αn if all players choose C in the first stage; αx if some players choose D in the first stage and $(s_i, t_i) = (C, S)$; and $\alpha x + 1$ otherwise. Let $s = (s_i)_{i \in I}$ and $t = (t_i)_{i \in I}$. Given s, let $u_i(t)$ be i's payoff when the second-stage decisions are t. Then, we obtain the following result.

Proposition 1. *Assume $n \geq 2$. If all players are selfish, and it is common knowledge, then for any $\alpha \in (1/n, 1)$, the unique BEWDS equilibrium outcome under the SLM is cooperative.*

Proof. Let $n \geq 2$ and $\alpha \in (1/n, 1)$. Suppose that all players are selfish. We first prove by induction that in every second-stage subgame, every C player chooses L on the number of C players k, where $k \in \{1, 2, ..., n-1\}$. Without loss of generality, assume that each $i \in \{1, 2, ..., k\}$ is a C player. It suffices to show that player 1 chooses C in the first stage. Consider $k = 1$: $s = (C, D, ..., D)$. Then, player 1 is the only player who proceeds to the second stage. Because choice L yields 1 whereas choice S yields $\alpha < 1$, L dominates S. The induction hypothesis is that if $s = (\underbrace{C,...,C}_{k}, D, ..., D)$, for any $(t_2, ..., t_k)$,

$$U_1(L) := u_1(\underbrace{L, t_2, ..., t_k}_{k}, -, ..., -) > u_1(\underbrace{S, t_2, ..., t_k}_{k}, -, ..., -) := U_1(S).$$

Consider $s = (\underbrace{C,...,C}_{k+1}, D, ..., D)$. By construction of the SLM, we have for any $(t_1, ..., t_k)$,

(1) $u_1(\underbrace{t_1, t_2, ..., t_k, S}_{k+1}, -, ..., -) = U_1(t_1) + \alpha$ and $u_1(\underbrace{t_1, t_2, ..., t_k, L}_{k+1}, -, ..., -) = U_1(t_1)$.

Consider first the case when player k+1 chooses S. The induction hypothesis and (1) imply

that for any $(t_2,...,t_k)$,

$$u_1(\underbrace{L,t_2,...,t_k,S}_{k+1},-,...,-) = U_1(L)+\alpha > U_1(S)+\alpha = u_1(\underbrace{S,t_2,...,t_k,S}_{k+1},-,...,-).$$

Similarly, when player k+1 chooses L, we have

$$u_1(\underbrace{L,t_2,...,t_k,L}_{k+1},-,...,-) = U_1(L) > U_1(S) = u_1(\underbrace{S,t_2,...,t_k,L}_{k+1},-,...,-).$$

Hence, the hypothesis holds for k+1. The same argument holds for every C player.

Consider the reduced normal form game. Because every C player chooses L using this argument, player 1 gets αn only if all players choose C in the first stage, otherwise player 1 gets $1 < \alpha n$. Therefore, C weakly dominates D. ∎

Let us introduce non-selfish motive which is frequently observed in experiments. We say a player is *conditionally cooperative* if (i) he chooses C in the first stage; and (ii) he chooses *Leave* in the second stage if the number of other C players is less than l, and chooses *Stay* otherwise, where $l \in \{1, 2, ..., n-1\}$. We assume that l is common among conditionally cooperative type. We begin with the extreme case, as a counterpart of Proposition 1.

Proposition 2. *Assume* $n \geq 2$. *If all players are conditionally cooperative, and it is common knowledge, then for any* $\alpha \in (1/n, 1)$, *the unique outcome under the SLM is cooperative.*

Proof. Straightfoward by definition. ∎

2.3. Behavioral heterogeneity cases

We next consider a mixed population where a positive number of both selfish and conditionally cooperative types exist. Let $m \in \{1, 2, ..., n-1\}$ be the number of selfish types among n players. For the sake of simplicity, l and m are common knowledge. In order to illustrate that m is crucial to induce the cooperative outcome under behavioral heterogeneity, consider the following example of $n = 3$ and $\alpha = 0.7$. Note that $l \in \{1,2\}$. When $l = 2$, a conditionally cooperative type chooses *Leave* in subgames *CCD* and *CDD*; hence, the problem reduces to the case when all players are selfish (Proposition 1). We are interested in the case of $l = 1$, which means that the conditionally cooperative type chooses *Stay* in subgame *CCD*, but *Leave* in subgame *CDD*.

Assume first m = 2: there exist two selfish types and one conditionally cooperative type. Since the conditionally cooperative type's choice depends only on the others' first-stage choices, let us focus on the selfish type, say player 1. Note that if the first-stage

choices are *CCC*, player 1 receives 21. Consider subgame *DCC*. Since player 1 knows that only one player, say player 3, is conditionally cooperative, he expects that player 2 will choose *Leave*, while player 3 will choose *Stay*. Then, player 1's payoff (multiplied by 10) in subgame *DCC* is $10+0.7\times 10=17$. Likewise, by choosing *Leave* in subgame *CDC*, player 1 receives 17. In subgame *DDC*, on the contrary, player 1 expects that player 3 chooses *Leave*; hence, player 1 receives 10. Now, consider player 1's payoffs in the reduced normal form game presented in the left panel of Table 3. Since (21,17) weakly dominates (17,10), player 1 chooses *C*.

		Player 2 (selfish)	
		C	*D*
Player 1 (selfish)	*C*	21,21,21	17,17,7
	D	17,17,7	10,10,10

When *m*=2: players 1 and 2 are selfish types

		Player 2 (conditionally cooperative)	
		C	*D*
Player 1 (selfish)	*C*	21,21,21	17,17,7,,
	D	24,14,14	10,10,10

When *m*=1: only player 1 is a selfish type

Table 3. Reduced normal form game under the SLM when at least player 3 is conditionally cooperative (hence player 3's *C* is given).

Assume next $m = 1$: only player 1 is selfish and the other players are conditionally cooperative. It is sufficient to consider subgame *DCC*. Since player 1 knows that both of the other players are conditionally cooperative types, he expects that players 2 and 3 will choose *Stay*. Then, player 1's payoff (multiplied by 10) in subgame *DCC* is $10+0.7\times 20=24$. Since 24>21, player 1 chooses *D*. The shaded cells in Table 3 indicate the difference in outcome depending on the number of conditional cooperators. The above example suggests that an increase in the number of conditional cooperators reduces the overall cooperation rate.

In accordance with the above example, we focus on the nontrivial cases $l \le n-2$, where selfish and conditionally cooperative types choose differently. Otherwise, we can just apply Proposition 1 as if all players are selfish types. From $l \ge 1$ and $l \le n-2$, we focus on $n \ge 3$. We find a condition such that a mixed population of selfish and conditionally cooperative types cooperates under the SLM as follows.

Proposition 3. *Assume* $n \ge 3$. *If there are m selfish types and n-m conditionally cooperative types with threshold* $l \le n-2$, *and these facts are common knowledge, then the unique equilibrium outcome under the SLM is cooperative if and only if* $\alpha \in [1/m, 1)$.

Proof. Let $n \geq 3$, $m \in \{1, 2, ..., n-1\}$, $l \in \{1, 2, ..., n-2\}$, and let $\alpha \in (1/n, 1)$. We show that all players choose *C* in the first stage if and only if $\alpha \in [1/m, 1)$. Since any conditionally cooperative player's choice depends only on the others' first-stage choices, it suffices to check the incentive of BEWDS players. Without loss of generality, assume player 1 is a selfish type. Owing to the symmetry of the model, it suffices to consider player 1. Note that by definition *n*-*m* conditionally cooperative players choose *C*.

Consider a subgame where no more than *m*-2 selfish players except for player 1 choose *C* in the first stage. Since the total number of *C* players except for player 1 is no more than *m*-2+(*n*-*m*)=*n*-2, every *C* player proceeds to the second stage regardless of 1's first-stage choice. From the construction of the SLM, *Leave* is better than *Stay*, and choosing *C* then *Leave* is indifferent to choosing *D* for player 1.

Consider a subgame where *m*-1 selfish players except for player 1 choose *C* in the first stage. Suppose player 1 chooses *D* in the first stage. Since the total number of *C* players is $n-1 > l$, all *n*-*m* conditionally cooperative types choose *Stay* in the second stage, while all the other *m*-1 selfish types choose *Leave*. Then, in subgame $D\underbrace{C...C}_{n-1}$, 1's payoff is $1+\alpha(n-m)$. Suppose player 1 chooses *C* in the first stage. Then, the game ends and player 1 receives αn. If $\alpha = 1/m$, because $1+\alpha(n-m) = \alpha n$, *C* and *D* are indifferent. If $\alpha \in (1/m, 1)$, from the above argument and $1+\alpha(n-m) < \alpha n$, *C* weakly dominates *D*. Otherwise, *D* weakly dominates *C*. ∎

Proposition 3 has two implications for games that have three or more players. The first implication is that as long as there is at least one conditional cooperator who allows some defectors, it is impossible to achieve the cooperative outcome for any social dilemma case (because there exists $\alpha \in (1/n, 1/m)$). The second implication is that there is a tradeoff between the number of conditional players and the range within which the SLM achieves the cooperative outcome.

Before proceeding to the experimental design, we refer to the approval mechanism (AM) introduced by Saijo et al. (2015). After choosing *C* or *D*, all subjects proceeded to the second stage and were asked to either *Approve* or *Disapprove* the other players' first-stage choices. If all group members chose *Approve*, the outcome was the one they chose in the first stage. Otherwise, the outcome was the one when all three group members chose *D*.[10] If $n = 2$, the AM achieves the cooperative outcome in BEWDS. For $n \geq 3$, however, the AM does not.

[10] We simply call this mechanism the AM, although Saijo et al. (2015) consider only the two-player case.

3. Experimental design

We conducted three treatments, the SLM, AM, and SD as a control, at Osaka University in October 2012 and in January and March 2013. We set $n = 3$ and $\alpha = 0.7$ so that theoretically both cooperative and uncooperative outcomes could occur. Remember the three-player example in Section 2.3. The cooperative outcome is achieved if two selfish types exist (or are expected to exist according to selfish types) in a group by $1/2 < 0.7$, but the cooperative outcome is not achieved if only one selfish type exists (or is expected to exist according to selfish types) in a group by $1 > 0.7$.

We use a random matching protocol. In every period, each subject was given 1000 experimental currency units (ECUs). That is, if all three group members choose *D*, they each get 1000. Each SLM and AM has three sessions, and the SD has two sessions. In each session, 21 subjects played the game for 15 periods, but the third session of the AM consisted of 18 subjects. No individual participated in more than one session. Subjects were recruited from Osaka University through campus-wide advertisements. We used the z-Tree software (Fischbacher, 2007).

Each subject was randomly seated at a computer terminal, all of which were separated by partitions. Communication was prohibited among subjects. Each subject received written instructions and record sheets (see supplementary materials). An experimenter read the instruction out loud, and subjects were then given 5 minutes to ask questions. Then, there was no practice period, and subjects proceeded to the payment periods. In each period, subjects were anonymously divided into groups of three. We used a random matching protocol. We informed the subjects of this procedure.

The SLM treatment continued as follows. In the first stage (called the choice stage in the experiment) of each period, observing the payoff matrix, each subject was asked to select either *C* or *D* (which were presented using 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 clicked 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*, the group members proceeded to the result screen explained later. Otherwise, each *C* player proceeded to the second stage. In the second stage, 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. On the other hand, *D* players could not proceed to the second stage, so 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 the subjects. After finishing all 15 periods, subjects were asked to complete a questionnaire, after which they were immediately and privately paid in cash. Each subject was paid an amount proportional to the sum of ECUs that he had earned over the 15 periods.

In addition to these tasks, subjects answered a hypothetical questionnaire at the beginning of each period regarding their choices and on what they think their group members' choices would be in the first-stage and second-stage subgames. Hereafter, we call these the *pre-period questionnaires* (see the Appendix for the complete list of pre-period questions). Although there are six second-stage subgames in total, owing to the symmetry of the other two players, it sufficed to ask about four subgames, namely *CCD*, *CDD*, *DCC*, and *DCD*, where the first character indicates the responder's own choice in the first stage. After they completed the questionnaire, subjects proceeded to the first stage.[11]

In the AM treatment, we also conducted the pre-period questionnaires to elicit the subjects' belief on how many group members would choose *Approve*. Finally, the SD treatment did not include a second stage.

4. Experimental results

4.1. Average cooperation rates

Figure 2 shows the time path of the average cooperation rate over the 15 periods sorted by mechanism. We use the cooperation rates after the second stage in the SLM and AM.

[11] Before the second stage, subjects also answered questionnaires asking what they would hypothetically choose and what they think *C* players would choose in the subgame the group actually reached. We did not find notable results for this questionnaire.

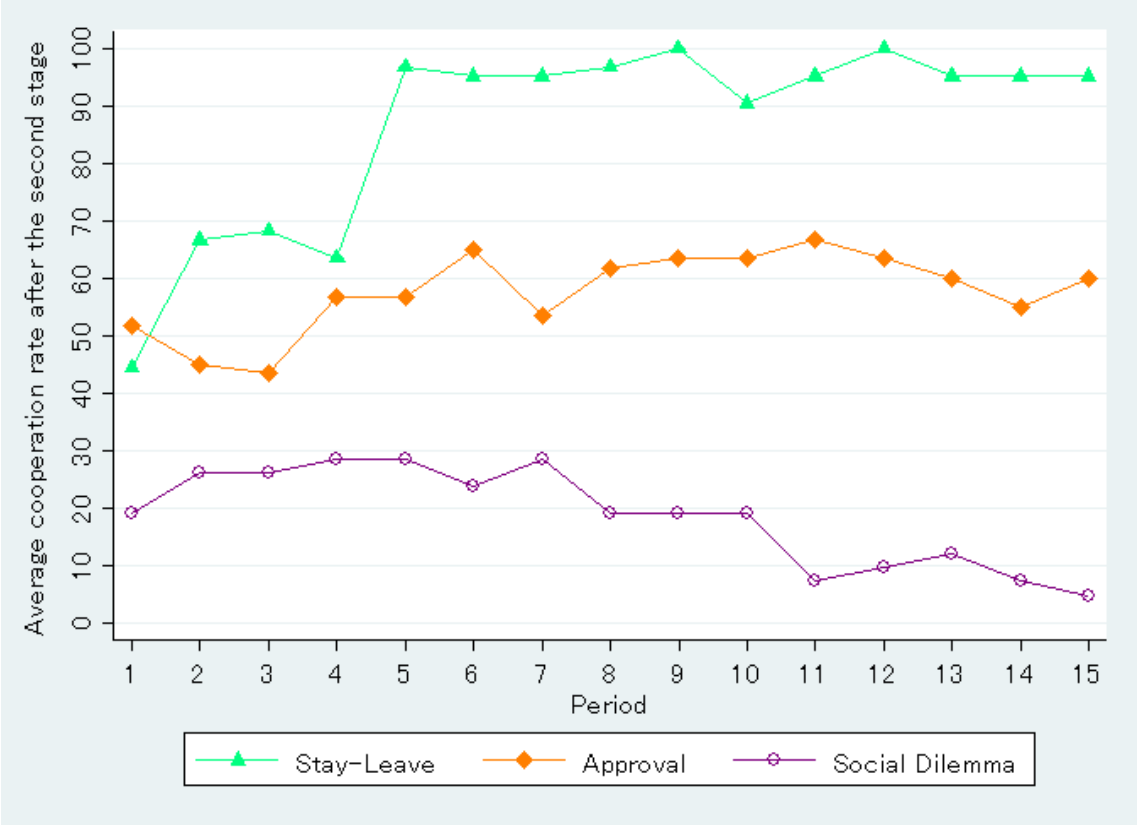

Figure 2. Average cooperation rate after the second stage by period and sorted by mechanism.

Result 1. *Subjects in the SLM cooperated on average 86.6% of the time. In periods 5–15, the average cooperation rate was 96.0%.*

Support. The average cooperation rate in the SLM sessions (the line with the triangle symbols) was 44.4% in the first period. After four periods, the average cooperation rate increased to at least 90%. 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. Spearman's rank correlation test supports the convergence to the cooperative outcome, showing that the upward time trend in the average cooperation rate under the SLM was statistically significant $(\rho = 0.5359, p = 0.0395)$.

A first look at the AM sessions, denoted by the line with diamond symbols, shows that the cooperation rates varied within the middle range during the experiment. The overall average cooperation rate of the AM was 57.7%. Although the upward time trend of the average cooperation rate under the AM was statistically significant (Spearman's rank

correlation test $\rho = 0.5224, p = 0.0457$), the average cooperation rate was still far from that under the SLM, even in the final period.

The SD sessions (the line with circle symbols) replicated the observed pattern of previous experimental studies of social dilemma. In the first period, subjects cooperated 19.0% 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 nine of the 210 groups achieved a cooperative outcome. The downward trend in the average cooperation rate was statistically significant (Spearman's rank correlation test; $\rho = -0.7899, p < 0.001$).

Table 4 compares the cooperation rates after the second stage for the three treatments, using a two-tailed Wilcoxon rank sum test. Subjects in the SLM cooperated significantly more than those in the AM (test statistic $z = 6.870$; $p < 0.001$). Similarly, the AM also facilitated cooperation compared with the SD ($z = 6.061$, $p < 0.001$).[12]

Treatment	AM	SD
SLM	6.870**	7.938**
AM		6.061**

Note: ** $p < 0.01$.

Table 4. Wilcoxon rank sum tests for equality in the average cooperation rates after the second stage by mechanism.

4.2. Second-stage behavior under the SLM

This subsection explores group behavior under the SLM by stage to explain the evolution of the average cooperation rate. We divide the data into two parts: periods 1–4, where subjects deviate from BEWDS, and periods 5–15.

Result 2. *In the second-stage subgames under the SLM, we observed the following.*
(i) In periods 1–4, all seven CDD groups ended up with DDD. Among the 29 CCD groups, one group ended up with CCD, 10 groups ended up with CDD, and 18 groups ended up with DDD.
(ii) In periods 5–15, the first-stage deviation from BEWDS was only CCD. Among the 10 CCD groups, two groups ended up with CDD and eight groups ended up with DDD.

Support. Table 5 tabulates the distribution of the group-level choices observed in periods 1–4 and 5–15. The rows indicate the first-stage choices and the columns indicate the final

[12] Following Andreoni and Miller (1993), we use the average cooperation rate by subject in order to eliminate serial correlation.

choices after the second stage. That is, if a player chooses (*C*, *S*) (resp. (*C*, *L*)), we denote the player's choice as *C* (resp. *D*) in the column.

		Periods 1–4					Periods 5–15					
		Final Choices					Final Choices					
		DDD	*CDD*	*CCD*	*CCC*	Total	*DDD*	*CDD*	*CCD*	*CCC*	Total
First-stage choices	*DDD*	1				1	0				0
	CDD	7	0			7	0	0			0
	CCD	18	10	1		29	8	2	0		10
	CCC				47	47				221	221
	Total	26	10	1	47	84	8	2	0	221	231

Notes: a) The shaded cells indicate that the corresponding outcome is not applicable. b) Thick-framed cells indicate the BEWDS prediction for the subgame of the corresponding row.

Table 5. The distribution of the group choices under the SLM.

In periods 1–4, the second-stage subgame outcomes seem to depend on the first-stage choices even though *DDD* is the unique BEWDS prediction in any second-stage subgame. If there was only one *C* player in the group, the player chose *Leave* in all cases, consistent with BEWDS. If there were two *C* players in the group, however, a deviation from BEWDS frequently occurred. *C* players chose to *Stay* in 37.9%(=(10+1)/29) of the groups facing *CCD*.

The responses to the pre-period and post-experiment questionnaires provide clues to why subjects sometimes chose *Stay* at *CCD*. We find that seven out of 14 *C* players chose *Stay* at *CCD* because they expected the other *C* player to also choose *Stay*. Nevertheless, such *C* players ended up with *CDD*.[13] Four subjects described in the post-experimental questionnaire on subgame *CCD* that (*Stay, Stay*) yields the cooperative outcome for two *C* players, (1400, 1400). One possible explanation of this observation is conditional cooperation.[14]

During periods 5–15, only *CCD* occurs as an observed first-stage deviation, and the two *C* players chose *Leave* 80.0%(=8/10) of the time. The proportion of second-stage choices

[13] There are six cases in periods 1–4 and one case in period 5.

[14] As shown in Table 5, the average *Stay* rate of *C* players for *CCD* is 17.9% = (10+1·2+2)/2·(29+10), which is higher, but not significantly so, than the average cooperation rates in the prisoner's dilemma experiment (overall average: 10.0%) with random matching in Saijo and Okano (2015; Wilcoxon rank sum test, $z = 1.192$, $p = 0.2331$).

consistent with BEWDS increased over the experimental periods (such outcomes occurred 69.4%(=(7+18)/(7+29)) of the time for *CDD* and *CCD* in periods 1–4).

4.3. Classification of subject types by using the pre-period questionnaire

In this subsection, we clarify subjects' behavioral types by checking the combination of the first-stage choices and answers to the pre-period questionnaires. We allow subjects to change their types over time. The following result suggests the coexistence of selfish and conditionally cooperative types throughout the session.

Result 3. *In the first period, 44.4% of subjects were deemed to be selfish types and 19.0% were deemed to be conditionally cooperative types. Type heterogeneity thus remains even after repetition, changing the type composition through a 34.8% increase in selfish types and a 9.0% decrease in conditionally cooperative types.*

Support. The rows in Table 6 summarize the share of behavioral types, broadly categorized into selfish, conditionally cooperative, other subjects choosing *C* in the first stage, and other subjects choosing *D* in the first stage. The rightmost three columns show the shares of these four types in period 1, where initial belief can be elicited, periods 1–4, where the cooperation rates remain low, and periods 5–15, where the cooperation rates remain high. The type-specific first-stage choices and beliefs are listed after the second column.

For example, consider a selfish type believing both of the other players are also selfish. As stated in Proposition 1, such a type, in the pre-period questionnaire, will explain that in every second-stage subgame, each player is expected to choose *Leave.* Then, thinking backwardly, this type chooses *C* in the first stage. Moreover, since this selfish type uses domination between strategies, there is no restriction on the answers given to explain the expected first-stage choices of the other players, as indicated by the "-" symbol. Let us take another example of a conditionally cooperative type. As we have seen in the three-player example before Proposition 3, this type chooses *C* in the first stage and *Stay* in subgame *CCD*, but chooses *Leave* in subgame *CDD*. Again, there is no further restriction on the answers given to explain the expected choices of the other players.

When we focus on the share in the first period, 44.4% (=28/63) of subjects are classified as selfish types, while 19.0% (=12/63) are deemed to be conditionally cooperative types. Since no one had experienced the game at this point, this evidence suggests that type heterogeneity is innate.

Behavioral Type (expectation of others' type)	First-stage choice	Answer for the expected first-stage choices of other players	Answer for own second-stage choice in subgame		Answer for the expected second-stage choices of other players in subgame			Percentage		
			CDD	*CCD*	*CCD*	*DCD*	*DCC*	Period 1	Periods 1-4	Periods 5-15
Selfish (both are also selfish)	*C*	-	*Leave*	*Leave*	*Leave*	*Leave*	*Leave, Leave*	34.9	50.8	75.6
Selfish (one is also selfish, but the remaining one is conditionally cooperative)	*C*	*CD or CC*	-	*Leave*	-	*Leave*	*Stay, Leave*	3.2	4.8	3.3
Selfish (both are conditionally cooperative)	*D*	*CC*	-	-	-	-	*Stay, Stay*	6.3	2.0	0.3
Selfish subtotal								44.4	57.6	79.2
Conditionally cooperative	*C*	-	*Leave*	*Stay*	-	-	-	19.0	15.1	10.0
Other *C*	*C*	-	-	-	-	-	-	12.7	11.1	9.7
Other *D*	*D*	-	-	-	-	-	-	23.8	16.3	1.2
(*C* by selfish or conditionally cooperative)/(*C* total)								81.8	86.4	90.2
Total								100.0	100.0	100.0

Notes: "-" indicates any answer is plausible.

Table 6. Classification of behavioral types in the SLM sessions according to the answers given to the pre-period questionnaire.

We also find the tendency for players to shift to become selfish as experience rises, although a proportion of subjects remain conditionally cooperative throughout. For example, while the selfish type amounts to 57.6% (=145/252) in periods 1–4, this increases to 79.2% (=549/693) in periods 5–15. On the contrary, conditional cooperators comprise 15.1% (=38/252) in periods 1–4, but his proportion decreases to 10.0% (=69/693) in periods 5–15. Moreover, selfish and conditionally cooperative types constantly explain over 80% of the first-stage *C*.

The data in Table 6 also suggest why the SLM works well after period 5. This success is partly attributed to the fact that most selfish types expect the other two players to also be selfish (approximated by $m = 3$). Such an expectation can be said to be roughly consistent with empirical composition of types. If selfish types had have kept misbelieving that a large proportion of the subjects were conditionally cooperative (approximated by $m = 1$), they would have chosen *D*.[1]

5. Concluding remarks

Previous laboratory- and field-based experimental studies of social dilemma have provided substantial evidence on the existence of behavioral heterogeneity among subjects as represented by selfish and conditionally cooperative types, even in games with symmetric payoffs. To avoid conditionally cooperative types being taken advantage of by selfish ones, several approaches have been proposed, such as introducing some game form (e.g., auction), the redistribution of intertemporal payoffs, endogenous type sorting, and so on.

In line with the first approach, we introduced the SLM for n-player prisoners' dilemma games to achieve cooperation among selfish and conditionally cooperative players. Under the SLM, each cooperator has the chance to revise his choice when players' choices are not unanimous. We considered two cases, namely when players have the same behavioral type and when they do not. The theoretical predictions between these two cases are contrasting. In the former case, the unique outcome is a cooperative one for any marginal value of cooperation, $\alpha \in (1/n, 1)$. In the latter case, where there are m selfish types, however, m and α must satisfy $\alpha \in [1/m, 1)$ for the uniqueness of the cooperative equilibrium outcome, as long as conditionally cooperative types allow the presence of a defector. The interpretation of this finding is straightforward. The gain from

[1] It should be noted that in total 41.7% (=46/107) of conditionally cooperative types tend to believe that the other two group members are also conditional cooperators, that is, they expect the other players to choose *C* in the first stage and *Leave* in subgame *DCD*, and for them both to choose *Stay* in subgame *DCC*. This observation supports our assumption that conditionally cooperative players choose *C* in the first stage.

switching to defection in the first stage, 1, does not exceed he total cost of the first-stage defection caused by his and by the remaining m-1 selfish players' switch to defection in the second stage, αm. Since the type composition endogenously emerges through players' interactions, we tested the SLM and type heterogeneity in the lab, using the parameters where both cooperative and uncooperative outcomes are theoretically predicted, depending on type composition.

In our experiment, we observed convergence to the cooperative outcome after period 5 with an average cooperation rate of 96.0%. This observation contrasts with the results of previous experimental studies of enhancing cooperation among homogeneously rational players in social dilemma settings. For example, in Varian's (1994) compensation mechanism experiment for two-player prisoner's dilemma games (Andreoni and Varian 1999; Charness et al. 2007), the cooperation rate remained at just 70%.

In order to explain why the SLM works well, we scrutinized the group-level outcomes and individual choices as well as players' responses to the questionnaire. These data supported the coexistence of selfish (44.4–79.2%) and conditionally cooperative types (10.0–19.0%) throughout the session. We also found the tendency to shift toward acting selfishly as experience rises. Moreover, selfish and conditionally cooperative types constantly explain over 80% of the first-stage choices to cooperate. Roughly consistent with the classification results, most selfish types expect the other two players to also be selfish, under which the SLM theoretically achieves the cooperative outcome. Therefore, our results corroborate the importance of incorporating behavioral heterogeneity into institutional design.

This study provides a way in which to explore game forms that achieve the cooperative outcome both theoretically and experimentally in a one-shot setting. Saijo and Okano (2015) modify the SLM to allow *D* players to revise their choices ahead of *C* players, observing that their mechanism yields a higher cooperation rate than the SLM does in earlier periods. Saijo and Masuda (2015) extend the SLM to a linear public good environment in order for players who have announced their maximum contribution to the group to revise their contributions freely. These authors show that in the latter half of the sessions with groups of five, namely periods 6–10, the average contribution rate is 95.3% under this mechanism. Future work could aim to extend these simple mechanisms to the general public good environment.

Appendix. Pre-period questionnaires

The questions in the pre-period questionnaires are listed as follows.

(1) Let's say you choose A and your two counterparts choose B in the choice stage. When

this happens, what do you think your two counterparts (who chose B and are advancing to the new choice stage) will choose?

Both of them will change to A

One will change to A and the other will stay with B

Both of them will stay with B

(2) Let's say you and one of the counterparts choose A and the other counterpart chooses B in the choice stage. When this happens, what do you think the counterpart (who chose B and is advancing to the new choice stage) chooses?

Change to A Stay with B

(3) Let's say you and one of the counterparts choose B and another counterpart chooses A in the choice stage. When this happens, what will you choose in the new choice stage?

Change to A Stay with B

In addition, what do you think the counterpart (who chose B and is advancing to the new choice stage) will choose?

Change to A Stay with B

(4) Let's say you choose B and your two counterparts choose A in the choice stage. When this happens, what will you choose in the new choice stage?

Change to A Stay with B

(5) Of the two counterparts, how many do you think will choose B in the choice stage?

0 1 2

Acknowledgments

We thank Takako Greve, Tetsuya Kawamura, Kazumi Shimizu, Masanori Takaoka, Takahiro Watanabe, and Hirofumi Yamamura. Keiko Takaoka also provided outstanding research assistance. This research was supported by JSPS KAKENHI Grant Number 24243028 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.

References

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., 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.

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

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

Bolton, G.E., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. American Economic Review 90, 166–193.

Brandts, J., Schram, A., 2001. Cooperation and noise in public goods experiments: applying the contribution function approach. Journal of Public Economics 79, 399–427.

Cason, T.N., Gangadharan, L., 2014. Promoting cooperation in nonlinear social dilemmas through peer punishment. Experimental Economics, February, 1–23.

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., Graziano, S., Maitra, P., 2006. Social learning and norms in a public goods experiment with inter-generational advice. The Review of Economic Studies 73(2), 357–380.

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.

Chen, Y., and Plott, C. R., 1996. The Groves-Ledyard mechanism: An experimental study of institutional design. Journal of Public Economics, 59(3), 335-364.

Croson, R., Shang, J., 2008. The impact of downward social information on contribution decisions. Experimental Economics 11, 221–233.

Dawes, R.M., 1980. Social dilemmas. Annual Review of Psychology 31(1), 169–193.

Falk, A., & Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior, 54(2), 293-315.

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

Gächter, S., 2007. Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications. CeDEx Discussion Paper no. 2006–03, University of Nottingham.

Grimm, V., Mengel, F., 2009. Cooperation in viscous populations-Experimental evidence. Games and Economic Behavior 66, 202-220.

Groves, T., Ledyard, J., 1977. Optimal allocation of public goods: A solution to the "free

rider" problem. Econometrica 45(4), 783–809.

Hamburger, H., 1973. N-person prisoner's dilemma. Journal of Mathematical Sociology 3(1), 27–48.

Healy, P. J., 2006. Learning dynamics for mechanism design: an experimental comparison of public goods mechanisms. Journal of Economic Theory, 129(1), 114–149.

Kalai, E., 1981. Preplay negotiations and the prisoner's dilemma. Mathematical Social Sciences 1(4), 375–379.

Kurzban, R., Houser, D., 2005. Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. Proceedings of the National Academy of Sciences of the United States of America 102(5), 1803–1807.

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

Ledyard, J. O., Palfrey, T. R., 1999. A characterization of interim efficiency with public goods. Econometrica 67(2), 435–448.

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

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.

Nowak, M. A., 2006. Five rules for the evolution of cooperation. Science, 314(5805), 1560–1563.

Ostrom, E., 2000. Collective action and the evolution of social norms. Journal of Economic Perspectives 14(3), 137–158.

Saijo, T., Masuda, T., 2015. The simplest solution to the free-rider problem: Theory and experiment. Unpublished manuscript.

Saijo, T., Okano, Y., 2015. Second thought: Theory and experiment in social dilemma. KUT-SDE working paper series no. 2014-7, Kochi University of Technology.

Saijo, T., Okano, Y., Yamakawa, T., 2015. The mate choice mechanism experiment: A solution to prisoner's dilemma. Unpublished manuscript.

Schelling, T.C., 1973. Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. Journal of Conflict Resolution 17(3), 381–428.

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