KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

Discussion Paper No.944

"Cooperation among Behaviorally Heterogeneous Players
in Social Dilemma with Stay or Leave Decisions"

Xiaochuan Huang, Takehito Masuda, Yoshitaka Okano, and Tatsuyoshi Saijo

July 2016

KYOTO UNIVERSITY

KYOTO, JAPAN

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

Xiaochuan Huang[a], Takehito Masuda[b,*], Yoshitaka Okano[c], and Tatsuyoshi Saijo[d]

July 2016

[a]*DT Capital Management Co., Ltd., North Tower, 35 Sinan Road, Shanghai, 200020, China*
huangxiaochuan7246@126.com
[b]*Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan.*
takehitomasuda@gmail.com
[c]*School of Economics and Management, Kochi University of Technology, 2-22 Eikokuji, Kochi, 780-0844, Japan*
okano.yoshitaka@kochi-tech.ac.jp
[d]*School of Economics and Management, Kochi University of Technology, 2-22 Eikokuji, Kochi, 780-0844, Japan*
tatsuyoshisaijo@gmail.com

Abstract

We experimentally test a two-stage mechanism called the *stay-leave mechanism* to achieve cooperation in *n*-player prisoner's dilemma situations. Under this mechanism, each cooperator has the chance to revise his choice when players' choices are not unanimous. We say a player is selfish if he eliminates dominated choices in each stage. If all participants of the stay-leave mechanism are selfish, for any value of public good benefit that arises, the unique equilibrium is unanimous cooperation. The average cooperation rate in the stay-leave mechanism experiment averaged 86.6% across 15 periods, with an upward trend, increasing to 96.0% after period 5. By examining earlier period data, we detected that selfish and conditionally cooperative subjects coexist at a proportion of approximately 3:1. Finally, we extended our model to incorporate a mixture of the observed two types and misbeliefs about others' types. Paradoxically, unanimous cooperation is less likely to occur as the number of conditionally cooperative players increase. The model also partially explains the observed upward trend in the cooperation rate in the stay-leave mechanism sessions.

JEL Classification Codes: C72; C72; D74; H41; P43
Keywords: social dilemma; experiment; conditional cooperator; behavioral heterogeneity

* Corresponding author

1. Introduction

A long-standing question in economics is how to foster cooperation when conflicts exist between individuals and collectives, as represented by public good provision. Studies of mechanisms that aim to enhance cooperation in public good provision have developed a theory by employing the model of homogeneously selfish players. However, few studies have designed well-behaved mechanisms with human subjects.

Several mechanism experiments designed to solve of social dilemma problems provide evidence that a subject's motivation for contribution/cooperation is not homogeneously selfish but rather heterogeneous. First, we look at Varian's (1994) compensation mechanism. When applied to prisoner's dilemma games, the mechanism induces two-stage games where players can offer transfers contingent on cooperation before choosing cooperation or defection. Although the mechanism has a strong theoretical property (i.e., that mutual cooperation is the subgame perfect Nash equilibrium), the experimental data of Andreoni and Varian (1999) showed that it took 20 or more repetitions for two-thirds of subjects to reach the equilibrium. Subsequently, Charness et al.'s (2007) study of the compensation mechanism detected that the preference for the equity of the final payoffs after transfers hinders reaching the equilibrium. Second, Levati and Neugebauer (2004) took the opposite approach, assuming that all players are conditionally cooperative. The authors conducted an English auction experiment applied to linear public good provision, since it would yield efficiency under the above assumption. Nevertheless, the lab auction failed to do that, and the data showed that subjects are a mixture of selfish and conditionally cooperative players.[1]

Thus, the unproven and fundamental challenge that naturally arises is to design a mechanism that aligns with heterogeneous behavioral rules. In this vein, Saijo et al. (2016) introduced the approval mechanism where players can cooperate if they eliminate weakly dominated actions in each stage. The authors showed empirically that the approval mechanism yielded significantly higher cooperation rates than the compensation mechanism, using prisoner's dilemma games. The authors then classified individual choices and detected that the driving force behind subjects' behavior is heterogeneous, including reciprocity and inequity aversion other than the strategic motivation on which they focused. Given these results, Masuda et al. (2014) extended the

[1] In Levati and Neugebauer (2004), every bidder standing at the clock/price p must contribute at least p, analogous to an English auction, to sell indivisible objects. Then, the argument about the failure of English auctions is intuitive. If an English auction determines each player's contribution, selfish players drop out at the low contribution level, and this triggers other conditionally cooperative players to drop out as well.

approval mechanism so that it works in the linear public good environment. They then proved the extended mechanism is robust in the sense that it aligns five different motivations with cooperation as long as matched players follow the same behavior. Masuda et al. (2014) also gave some empirical support for behavioral heterogeneity.

The aim of the present paper is to enhance cooperation in the *n*-player prisoner's dilemma environment, taking the approach of Saijo et al. (2016) and Masuda et al. (2014). It introduces a two-stage mechanism called the *stay-leave mechanism* (SLM), where every player can either contribute (*C*, after cooperation) his endowment or not contribute at all (*D*, after defection). The SLM 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 *Stay* or *Leave*. If a player chooses *Stay*, he contributes the endowment. If the player chooses *Leave*, he makes no contribution. No *D* player proceeds to the second stage and thus contributes nothing.

The prediction that motivates our experiment is as follows. Assume that every player eliminates weakly dominated actions in each stage and that this is common knowledge. Then, for any marginal per-capita return such that the dilemma arises, the unique equilibrium outcome under the SLM is unanimous cooperation. In other words, the SLM implements unanimous cooperation when all players are selfish in the above sense.[2]

To test this prediction, we conducted experiments with groups of three and random matching. We find that introducing the SLM significantly increases the average final cooperation rates compared with *n*-player prisoner's dilemma only. In the SLM sessions using the direct method, 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. On the contrary, *n*-player prisoner's dilemma only sessions yield an average cooperation rate of only 18.5%. However, our original theory cannot explain the observed upward trends in SLM

[2] We believe that the novelty of the current paper mainly lies in the analysis of mixed behavioral types. We found the SLM, independently from Gerber et al. (2013), who examined a coalition formation game varying the minimum required number of players to make coalition. In game called IF*m*, each of *n* players decide to join or not to coalition for full contribution, and if there are *m* or more yes votes, such voters make full contributions immediately, while other no voters can freely choose contributions. By setting *m*=*n,* we have the game like the SLM, where every player plays voluntary contribution mechanism whenever decisions are not unanimous. On the other hand, we started from different question: can we find a mechanism with unique efficient equilibrium within a class of mechanisms with some stages *after choosing contributions* (discrete or continuous) to finalize them (Saijo et al., 2012 and Masuda et al., 2014). We can easily show that, in *n*-player prisoner's dilemma environment, using unanimity is required to achieve efficiency in the unique equilibrium of selfish players. Hence, the SLM is extension of approval mechanism in Saijo et al. (2012) to *n*-players but at the same time we can regard it as a simplified version of IF*n*, after changing the labels of choices (from *C, D, Stay, Leave* to *join*, *don't join, C, D*, respectively*).*

cooperation rates.

We find evidence of behavioral heterogeneity behind this upward trend in the cooperation rate. We thus conducted additional SLM sessions by using the strategy method to elicit subjects' choices in relevant but unrealized paths and obtain precise behavioral-type information. The behavioral-type classification using data derived from the strategy method shows that 50.0% of subjects are selfish and 14.7% are conditionally cooperative. Roughly, conditionally cooperative subjects are those who "cooperate if there is sufficient chance that their opponent will do likewise" (Andreoni and Samuelson, 2006). Our way of classification is simple: selfish subjects will choose *Leave* in every second stage subgame; on the contrary, conditionally cooperative subjects will choose *Stay* as long as they observe that other subjects in a group chose *C* in the first stage, expecting that the other subject will also choose *Stay* in the second stage.

Given these observations, we provide a novel model incorporating behavioral heterogeneity and incomplete information. Each selfish player has beliefs about the true number of conditionally cooperative players among n players. On the other hand, conditionally cooperative players initially believe that another *C* player will choose *Stay,* although selfish players never do that. This formulation is, although ad hoc, useful to represent the situation that each behavioral type may overestimate its own type among all players. First, we obtain a paradoxical result: when the support of beliefs positively correlates with true information, unanimous cooperation is less likely to occur as the number of conditionally cooperative players increase.[3]

Nevertheless, we can recover the unanimous cooperation result if we focus on repeated interaction rather than one-shot interaction. In fact, in our model unanimous cooperation occurs with a delay of one period. This finding explains the upward trend of cooperation observed in our experiment. An intuitive explanation of this delay is as follows. (i) The selfish player's overestimation of the number of conditionally cooperative players causes the selfish player's trial of exploitation; however, (ii) the beliefs about types are updated correctly since players' behavior reveals each player's type after one period of play; then, (iii) the selfish player knows that she can no longer exploit conditionally cooperative players.

At least two strands of research relate to the current paper, other than the literature on the mechanisms used to solve social dilemma problems above mentioned. First, this study contributes to experimental evidence on conditionally cooperative players. Our observation that a proportion of subjects are conditionally cooperative corroborates a stylized fact in the public good experiment literature (see Chaudhuri, 2011; Arifovic and

[3] We thank the reviewers for their encouragement on the extension of the model.

Ledyard, 2011). Among the several formulations proposed, we especially consider conditional cooperators to be driven by an "ex-ante belief about the contributions to be made by their peer," as expressed in Chaudhuri (2011, p. 56). Such a principled player approach is appropriate when contributions are simultaneous. Another formulation is to modify the utility function so that players' contribution choices affect each other. A recent sophisticated model is Steiger and Zultan (2014). The authors considered the sequential binary contribution game where players move one by one and a player's mental cost of cooperation reduces to some degree only if he observed that all previous movers cooperated.[4]

The growing endogenous coalition formation literature shares with us a research question asking how introducing social preferences affects the performance of institutions (Kosfeld et al., 2009; Gerber et al., 2013; Kube et al.; 2015). The endogenous coalition formation literature focuses on incentives outside mechanisms (before participation). On the other hand, the standard mechanism literature, including the current paper, focuses on incentives inside mechanisms (after participation).[5] Kosfeld et al. (2009) provided a positive answer to the question of whether social preference promotes efficiency. The authors considered a three-stage linear public good game where players form coalitions to activate centralized punishment applied only to coalition members. Observing that subjects are more likely to form a grand coalition than the selfish player model predicts, they concluded that the data suggest a large degree of inequality aversion.[6] More recently, Kube et al. (2015) concluded the opposite. The authors experimentally showed that social preference for equity may hinder efficiency. Broadly speaking, the current paper gives a similar message to Kube et al. (2015), although taking different approaches.

In light of the literature mentioned above, our contributions are threefold. First, we show that the SLM in the spirit of Saijo et al. (2016) behaves well with human subjects in the simple n-player environment. Together with Saijo et al. (2016), this finding suggests that the SLM would work better than Varian's compensation mechanism in n-player prisoner's dilemma games. This observation is consistent with Gerber et al. (2013), who

[4] See Arifovic and Ledyard (2011) for the different formulations of conditional cooperators in laboratory studies. We only refer to some of them: distributional concern (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000); mechanics to behave reciprocally (Charness and Rabin, 2002); and conditional cooperation as a result of strategic thinking in repeated games (Andreoni and Samuelson, 2006). See Gächter (2007) for field evidence about conditional cooperators.

[5] This is also called the institution formation literature. Although we can use the word institution and mechanism interchangeably, the literature often keeps the mechanism or institution applied among coalition members as a black box. Contributions of the members are somehow fixed. Instead, the literature often focuses on the effect of the minimum required number of members to social welfare.

[6] Using a three-stage game similar to Kosfeld et al. (2009), Dannenberg (2012) did experiment in situations where only small coalition and hence inefficiency is predicted. On the other hand, our motivation of experiment is to test mechanisms to achieve efficiency in the unique equilibrium.

reported unanimous voting rule worked better than the other majority rules in their linear public good experiment, motivated by Kosfeld et al. (2009), with little argument and evidence of social preference.

Second, we provide concrete evidence that the participants in our mechanism are clearly a mixture of selfish and conditionally types, by successfully combining the strategy method and questionnaire for belief elicitation about other players' choices in in mechanism sessions. Moreover, we also confirm that the observed behavioral type distribution is not significantly different between elicitation methods. The coexistence of selfish subjects and non-negligible proportion of subjects with social preference observations is consistent not only with studies to classify them such as Chaudhuri (2011) and Andreoni and Samuelson (2006), but also with mechanism experiment studies such as Andreoni and Varian (1999), Charness et al. (2007), and Levati and Neugebauer (2004).

Third, by developing a novel model incorporating two behavioral types and heterogeneous beliefs, we also show that increasing cooperativeness among players may paradoxically hinder efficiency theoretically. This message is in contrast to Kosfeld et al. (2009), rather in line with Kube et al. (2015). Nevertheless, our model has distinctive features from the one in Kube et al. (2015) based on Fehr and Schmidt (1999). First, our model of conditionally cooperative players is simply based on what subjects chose, and do not require preference parameters on equity. Second, by doing so, we can summarize the uncertainty on behvioral types into the number of conditionally cooperative players. Importantly, we show that behavioral heterogeneity can be a factor to hinder efficiency, in a paradoxical way, even without tension between efficiency and equit due to heterogeneity of marginal benefit of the public good, on which Kube et al. (2015)'s message relies.[7] Together with the second contribution, our paper thus suggests the importance of mechanism design with heterogeneous behavioral rules.

Readers may still have questions about the assumption underlying the SLM: the game stops under the SLM only if all players choose *C*. To explain this, it seems better to consider the SLM in the context of fundraising. Then, we can regard the SLM as a combination of respecting for unanimity and refund upon request if unanimity fails.[8]

---

[7] Although both Kube et al. (2015) and Kesternich et al. (2014) focused on interaction between heterogeneity in the value of public good and social preference, neither seem to estimate behavioral type composition from individual choice data.

[8] The argument of refunds and provision point, a cost to cover public good provision, is often applied to discrete good context (see Croson and Marks, 2000 for a meta-analysis). In the literature, contributions are often non-refundable if their total amount exceeds provision point, which is usually lower than the social endowment. The contributions blow threshold is automatically and fully refunded. As we referred, a similar assumption to respect for unanimity is commonly found in coalition formation games (Gerber et al., 2013, Kube et al. 2015). Even if we apart from unanimity, it is not uncommon that the endpoint of dynamic game for public good provision varies with what players chose (see Levati and

First, in this context, it is appropriate to assume that there is no designer who can simply force all players to choose donate (C) at the beginning, since the designer may be a mediator or target of donation.[9] Moreover, the rapidly growing collective donation movement called Giving Circles provides the evidence of unanimity rule in fundraising. Bearman, J. (2007) points out that 38 percent of 145 surveyed circles requires consensus of all members to finalize whether to contribute their pooled money mainly to community project, such as education, youth development, health and nutrition. The author also says that the number of members in a circle varies from 5 to 100, with diverse backgrounds, and small groups such as One Percent for Moms in New York tend to use consensus rule. Moreover, many circles employ identical contribution for every member. For these reasons, our model and mechanisms can well capture this situation. As of 2007, more than 400 Giving Circles exist in the US, and have raised more than $95 million (see also Eikenberry et al., 2009). The movement spread to Asia, especially in Singapore, India, China, Japan and so on.[10]

The remainder of this paper is organized as follows. Section 2 provides our model and prediction. Section 3 describes the experimental design. Section 4 discusses the experimental results and detects conditionally cooperative players. Section 5 describes the incomplete information model with a mixed behavioral type. Section 6 concludes.

2. The model

2.1. The Stay-Leave mechanism

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

Neugebauer, 2004 for English auction; Fréchette aet al. (2012) for legislative bargaining).

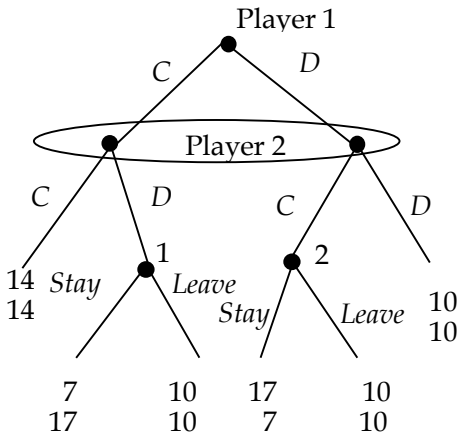
[9] It is highly necessary to make good fundraising mechanisms in life science, since even distinguished researchers do not have a coercive power. The iPS cell research group led by Dr. Shinya Yamanaka, who won the Nobel Prize for Physiology or Medicine in 2012, has been calling for contributions to the research fund publicly. It is clear that the huge benefits to life expectancy will be provided by the progress in iPS cell research, but the research group cannot force donations, while they may have to refund some donations because of the withdrawal of offers by donators. The SLM would be compatible with such a situation. For details, see https://www.cira.kyoto-u.ac.jp/e/about/fund.html.

[10] We thank Fumitaka Watanabe, who introduced us to Giving Circle.

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 a cooperative one (14, 14), that is, the Stay-Leave mechanism (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*),[11] 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, $(7,17)$. On the contrary, if player 1 chooses *Leave* at *CD*, the outcome is that when both defect, (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.

2.2. Theoretical prediction

In this subsection, we show that all players cooperate in the unique equilibrium. Our equilibrium concept come from Saijo et al. (2016), who developed a modified two-stage prisoner's dilemma game so that mutual cooperation is the unique equilibrium under the elimination of weakly dominated actions in each stage. In their experiment with perfect

[11] 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.

stranger matching, the authors observed a cooperation rate of 93.2%. Saijo et al. (2016) also showed experimentally that the above behavioral rule provides a clear prediction compared with the Nash equilibrium and subgame perfect Nash equilibrium. Masuda et al. (2014) designed a public good mechanism based on the same behavioral rule and experimentally verified that it works well. Therefore, the current paper continues to use the elimination of weakly dominated actions in each stage as the convincing behavioral rule. In what follows, we say that a player is selfish if he eliminates weakly dominated actions in each stage.

Next, we solve the game presented in Figure 1, assuming that all players are selfish. 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 thus construct the reduced normal form game shown in Table 1. Then, the pair of payoff player 1 may obtain by choosing *C* is [14, 10], while that 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).[12]

		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.

Before stating the first proposition, let us formulate 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, αkw, where $1/n < \alpha < 1$. In addition, each *D* player also receives the benefit from private consumption. Then, the total payoff is maximized when all players choose *C*, yielding $(\alpha nw, ..., \alpha nw)$, 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 his 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,

[12] We do not insist that backward elimination of weakly dominated actions best fits subjects' behavior in general class of games. Under the SLM, iterative elimination of weakly dominated actions 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*. Given that multiple interpretations are possible, we just say such players as selfish.

players simultaneously and privately choose *C* or *D*. If all 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 *Stay* or *Leave*. If the *C* player chooses *Stay*, he finally contributes *w*. If the *C* player chooses *Leave*, he contributes nothing. No *D* player proceeds to the second stage and thus contributes nothing. Now, we get the following result.

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

Proof. See Appendix.

3. Experimental design

We conducted experiments of the SLM sessions and SD sessions as a control, at Osaka University in October 2012, January and March 2013, and March 2016. Readers can refer to the online supplementary information to see all experimental materials. First, we explain the basic design across treatments.

Basic design across treatments

We set $n = 3$ and $\alpha = 0.7$ and 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. 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 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. In each period, subjects were anonymously divided into groups of three. We informed the subjects of random matching. 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.

Table 2 summarizes the experimental design. We vary behavior and belief elicitation method and payment scheme in SLM sessions.

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), c)]	6 (87)
SD	-	Total	2 (42)

Notes. a) For the list of questions, see Section 1 of Supplementary material. All pre-/mid-play questionnaires on others' actions are non-incentivized. b) A single period for payment is chosen manually by the experimenter, using number cards (1–15) and a box. c) The data from the SLM-strategy method sessions of the different two payment schemes are merged since there is no significant difference in the average first-stage cooperation rates between them.

Table 2. Summary of the experimental design.

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 by 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 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*, group members proceeded to the result screen 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. On the contrary, *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 subjects. There was no practice period in SLM-direct.

Prior to these tasks, subjects answered non-incentivized *pre-play questionnaires* 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. 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 (For the list of questions, see Section 1 of Supplementary material).[13] Finally, the SD treatment did not include a second stage.

SLM sessions by using the strategy method (SLM-strategy)

In order to check the robustness of the results obtained under SLM-direct, we additionally conducted the following treatments in March 2016. The first-stage is the same as in SLM-direct. Then, each subject who chose C in the first stage in SLM-strategy responded to questions asking his or her choice plan in both subgames *CDD* and *CCD*, before knowing what other group members chose in the first stage. Moreover, every subject answered, regardless of his or her first-stage choice, non-incentivized *mid-play questionnaires* set to elicit his or her belief about others' choices in the first stage and relevant second-stage subgames. Relevant subgames are *CDD* and *CCD* for each *C* player, while they are subgames are *DCD* and *DCC* for each *D* player (For the list of questions, see Section 1 of Supplementary material). Once other group members' planned actions are revealed, the relevant choices were realized. We also set a period of practice before the actual experiment to help subjects understand the rules, which seemed less intuitive compared with SLM-direct, as well as to minimize any effect on the payment periods.

SLM sessions with the strategy method and payment for a single period (merged to SLM-strategy)

We conduced six SLM-strategy sessions, which consisted of two different payment schemes, total and single period, each with three sessions.[14] For single-period payment sessions, the experimenter chose a number card manually from a box, in front of the subjects. Since there is no significant difference in the first-stage cooperation rates between payment schemes, we merged the data on both payment schemes.[15,16]

---

[13] 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.

[14] We thank the anonymous referee who suggested sessions with a single period payment.

[15] Blanco et al. (2014) reported a similar non-significant impact of payment schemes in social dilemma experiments.

[16] We omitted the sessions of a simple extension of approval mechanism in Saijo et al. (2016), since it does not achieve efficiency in theory. The data can be found in the working paper version. Similarly, Fischer, S., & Nicklisch, A. (2007) added unanimity voting stage after voluntary contribution stage in

4. Experimental results and feedback on model extensions

4.1. 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 second-stage cooperation rates for simplicity) in SLM-direct and SLM-strategy.

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

When we refer to second-stage cooperation rates, we also exclude the data of subjects who answered, in the post-experimental questionnaire, that they had not understood the rule of the experiment until period 7, the middle point of the session.

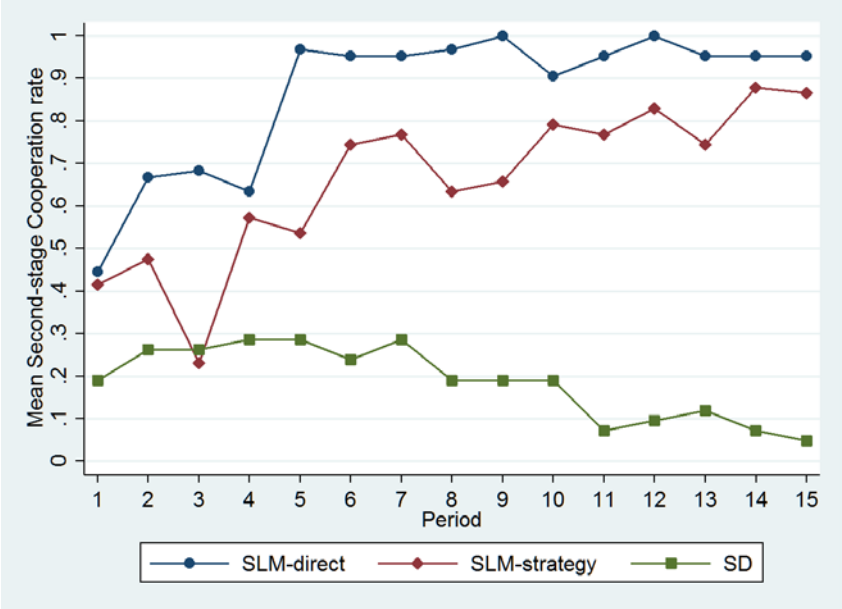


Figure 2. Average cooperation rate after the second stage by period and sorted by mechanism.

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. If we focus on the time after period 5, the average cooperation rate increased to at least

linear public good environment, but it also failed to yield efficiency.

96.0%. In fact, 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 $(p < 0.001)$.

The SD sessions (the line with the square symbols) replicated the observed pattern of previous experimental studies of SD. 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; $p < 0.001$).

As for the SLM-strategy sessions (the line with the diamond symbols), in period 1 the second-stage cooperation rate averaged 41.4%, similar to the one of SLM-direct. Across 15 periods, the second-stage cooperation rate averaged 65.0%. Spearman's rank correlation test supports the increasing time trend to the cooperative outcome $(p < 0.001)$.

Result 2. *SLM significantly increases average cooperation rate compared to the SD. Nevertheless, strategy method had significantly negative impact on average cooperation rate.*

Table 3 reports the marginal effects in a probit regression to compare the treatments in terms of the average cooperation rate, with standard errors clustered by session. The dependent variable is the second-stage cooperation (1) or defection (0) for each subject and each period. As regressors, we use three dummy variables*: D_SLM*, *D_strategy*, and *D_paysingle*, each respectively indicating that subjects participate in the SLM, in the strategy method sessions, and in the single period payment sessions, as well as *Period*. Specification (1) in Table 3 shows that the marginal effect of *D_SLM* is 0.497 and significant (at $p<0.001$), whereas the marginal effect of *D_strategy* is -0.256 and significant (at $p<0.01$).[17] Specification (1) also supports the upward trend of the cooperation rate, indicating that the marginal effect of *Period* is 0.0277 and significant (at $p<0.01$). These results qualitatively do not change when we use specification (2), including *D_paysingle.* This observation supports that we merged the data of all strategy method sessions.

t[17] Not all the Mann–Whitney test results are consistent with those obtained from the regression analysis. When we perform the Mann–Whitney test, we follow Andreoni and Miller (1993) and Charness et al.'s (2007) 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 Wilcoxon–Mann–Whitney tests, with grouping by treatment. We find that introducing the SLM significantly increases the average second-stage cooperation rate (*p*-values of the two-sided Mann–Whitney test for SLM-direct vs. SD < 0.001). This holds regardless of the elicitation method (*p*-values of the two-sided Mann–Whitney test for SLM-strategy vs. SD <0.001). There is also no significant difference between the two sessions (*p*-values of the two-sided Mann–Whitney test for SLM-direct vs. SLM- strategy =0.2121).

	Dependent Variable: final choice: *C* (and *Stay*)=1, *D* (or *Leave*)=0	
	(1)	(2)
D_SLM	0.497***	0.568***
	(0.0805)	(0.0805)
D_Strategy	-0.256**	-0.182***
	(0.0818)	(0.0390)
Period	0.0277**	0.0281**
	(0.0105)	(0.0107)
D_paysingle		-0.162
		(0.1480)
N	2805	2805
pseudo R^2	0.251	0.259

Regressors: *D_SLM:* dummy variable to indicate that subjects participate in *SLM*. *D_strategy:* dummy variable to indicate that subjects participate in strategy method sessions. *D_paysingle:* dummy variable to indicate that subjects participate in single period payment sessions. *Period:* variable that counts payment periods. Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Probit regressions (marginal effects).

4.2. Second-stage behavior under SLM-direct

This subsection explores group behavior under SLM-direct by stage to explain the evolution of the average cooperation rate. Table 4 tabulates the distribution of the group-level choices observed in periods 1–4. Note that the data shown in Table 4 come from incentivized choices in the sequence of the game, not from the data from the pre-play questionnaire. The rows indicate the first-stage choices and the columns indicate the final choices after the second stage. That is, if a player chooses *C* then *Stay* (resp. *C* then *Leave*), we denote the player's choice as *C* (resp. *D*) in the column.

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

Notes: a) The shaded cells indicate that the corresponding outcome is not applicable. b) Thick-framed cells indicate the choices selfish players will make for the subgame of the corresponding row.

Table 4. The distribution of the group choices under SLM-direct.

In periods 1–4, the second-stage subgame outcomes seem to depend on the first-stage choices even though *DDD* is the unique prediction of selfish players in any second-stage subgame. If there was only one *C* player in the group, the player chose *Leave* in all cases, consistent with the selfish model. If two *C* players were in the group, however, a deviation from selfish frequently occurred. The overall proportion of players who chose Stay in the *CCD* group was 20.7% (=(10+1*2)/29*2).

The responses to the pre-period and post-experimental 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*. Four subjects described in the post-experimental questionnaire on subgame *CCD* that both choosing *Stay* yields the cooperative outcome for two *C* players, (1400, 1400). One possible explanation of this observation is conditional cooperation.

4.3. Behavioral type classification by subject

In this subsection, we clarify subjects' behavioral types by checking the combination of the first-stage choices and answers to the questionnaires about the second-stage choices. Note that in what follows, we focus on the period 1 data in SLM-strategy, since they are well incentivized and independent among subjects. Another advantage of focusing on the first period data for the classification is that it simplifies matching subjects with their behavioral type. Table 5 shows how subjects were classified by behavioral type.

Behavioral Type	First-stage choice	Answer for own second-stage choice in subgame		Expectation for second-stage choices of other players in subgame *CCD*	Expectation for second-stage choices of other players in subgame *DCC*	Answer for the expected first-stage choices of other players
		CDD	*CCD*			
Selfish *C*	*C*	*Leave*	*Leave*	*Leave*	-	-
Conditionally cooperative	*C*	*Leave*	*Stay*	*Stay*	-	-
Selfish *D*	*D*	-	-	-	*Stay, Stay*	*CC*

Table 5. Classification criteria for behavioral types.

First, we classify a subject as *selfish C* if she satisfies the following conditions:

a) she chose *C* in the first stage;

b) she chose *Leave* in subgame *CDD*;

c) she chose *Leave* in subgame *CCD*; and

d) she expected that another *C* player would also choose *Leave*

Note that, as we explained in Section 3, we did not allow the *C* player's expectation of others' choices in subgames *DCD* and *DCC* in SLM-strategy to avoid overly complicating the task design. Hence, we use the above criteria, although they are insufficient to obtain a reduced normal form game (see Table 1), which requires the expectation that others choose *Leave* in subgames *DCD* and *DCC*.

Second, we classify a subject as *conditionally cooperative* if she satisfies conditions a), b),

e) she chose *Stay* in subgame *CCD*; and

f) she expected that another *C* player would also choose *Stay*;

Note that our classification might cause the underestimation of conditionally cooperative subjects, who are classified as selfish. This is because negatively reciprocal subjects, namely those who chose *Leave* because they did not expect other *C* players to choose *Stay* initially, are classified as selfish.

Third, we classify a player as *selfish D* if she satisfies the following conditions:

g) she chose *D* in the first stage;

h) she expected both *C* players to choose *Stay* in subgame *DCC*; and

i) she expected both of the other group members to choose *C* in the first stage.

Then, we find the following classification result.

Result 3. *Subjects in the SLM-strategy sessions exhibit behavioral heterogeneity. In particular, selfish C subjects and conditionally cooperative subjects account for about two-thirds of the data in period 1. Moreover, the behavioral type distribution is not significantly different between elicitation methods.*

Figure 3 is a horizontal bar graph showing the percentages of each behavioral type in the first period of SLM-strategy and SLM-direct. Each colored area in the stacked bar chart represents subjects classified as selfish *C*, conditionally cooperative, or selfish *D*.eusei

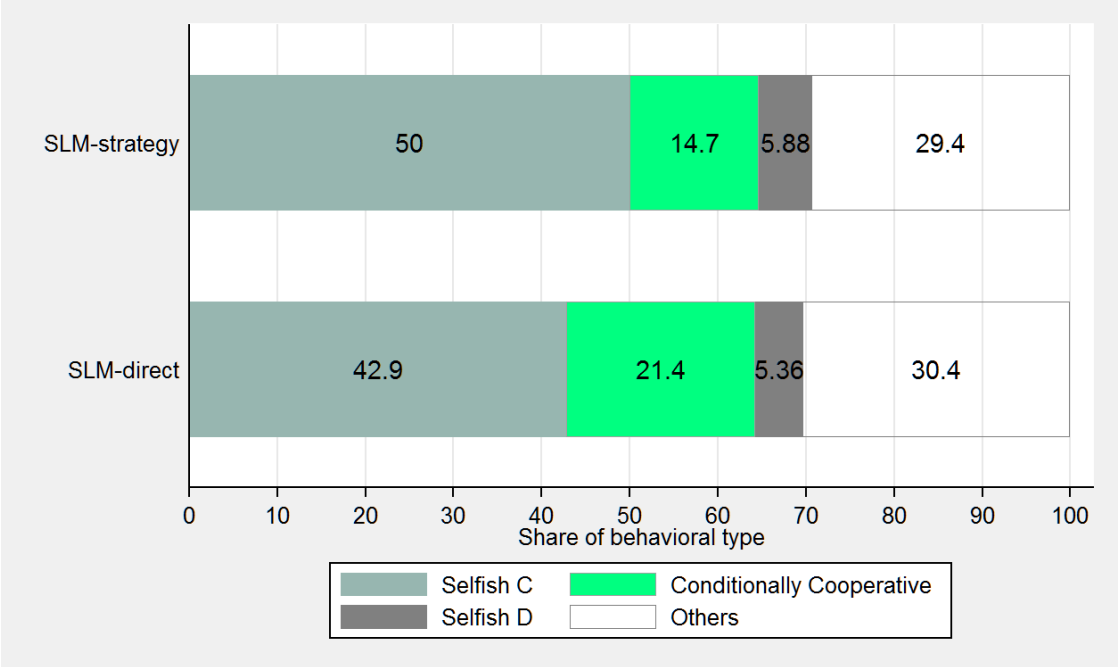


Figure 3. Percentages of behavioral type in period 1.

We restrict our attention to data from subjects who said they understood the rules at the beginning of period 1 in the questionnaires completed after the experiment. When limiting the data in this way, the total number of observations (number of participants) for SLM-direct and SLM-strategy were 56 and 68, respectively.

Interestingly, the results show that subjects were not homogeneous in the principles governing their behavior. Two typical behavioral types explain as much as two-thirds of the data, with 34 (50.0%) of the 68 subjects in the first period being selfish and 10 (14.7%)

being conditionally cooperative. In addition, 4 (5.9%) subjects were selfish *D*. Finally, as suggested in Figure 3, we confirm the robustness of behavioral heterogeneity. That is, the observed type distribution in SLM-strategy is not significantly different from that in SLM-direct (p = 0.767, chi-squared test).[18]

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

In this section, we extend the model so that players are a mixture of selfish and conditionally cooperative ones. They only know that there are two types, but do not know the exact number of each type. Facing such uncertainty, each player has a subjective belief about the number of conditionally cooperative players. We first derive the paradoxical result that *unanimous cooperation is less likely to occur as the number of conditionally cooperative players increase,* assuming one-shot interaction. Second, we show that *unanimous cooperation may be achieved afterward*, via belief updating under repeated interaction. Both predictions are consistent with our experimental data, which exhibit an upward trend of the average cooperation rate in earlier periods.

To define conditionally cooperative types by strategy rather than by their observed choices, we slightly modify the conditions used in the data classification. That is, the type chooses *Stay as long as* she believes the other *C* player will choose *Stay*. We assume the type initially believes that at least one other *C* player will choose *Stay*.

Example 1. Consider the example of $n = 3, \alpha = 0.7,$ and $w = 10.$ Suppose that players 1 and 2 are selfish and player 3 is conditionally cooperative. Prior to period 1, players have initial beliefs about the type composition of the other two players as follows. Player 1 (resp., 2 and 3) believes that two (resp., zero and one) of the other players will be conditionally cooperative. For simplicity, we write such a combination of beliefs as (2,0,1). Note that no player has the correct beliefs initially, since the correct belief profile is (1,1,0).

Table 5 shows the subjective reduced normal form game for player 1. We omit player 3's *C* since it is obvious. Light gray indicates that player 1 need not consider as long as she believes both of 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 subjective expected 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. Next, consider player 2. Since player 2 expects that both of the other players are selfish, the problem reduces to Proposition 1. Hence, player 2 chooses *C* (then *Leave*). Finally, player 3 chooses C and then *Stay*, since he

[18] To classify subjects in SLM-direct, we applied the conditions in Table 5 to the first-stage choices and the responses to the pre-play questionnaire.

is conditionally cooperative defined as above. Therefore, (*D*, *C* then *Leave*, *C* then *Stay*) occurs, yielding the payoffs of (17,17,7).

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

Table 6. Player 1's subjective reduced normal form game in period 1.

The above examples suggest that with heterogeneous beliefs, selfish players' overestimation of the number of conditionally cooperative players hinders cooperation. Then, we formalize this intuition by using the *n*-player model. For *n*-players, a player is *conditionally cooperative* if she satisfies the following conditions:

j) she chooses *C* in the first stage;

k) she chooses *Leave* in subgame *CD…DD*; and

l) she chooses *Stay* in any other subgame except for *CD…DD as long as* she expects another *C* player to also choose *Stay*.

m) she initially believes that at least one other *C* player will choose *Stay*.

Assume that $n \geq 3$ players consisting of *c*, $1 \leq c \leq n-1$ conditionally cooperative players and $(n - c)$ selfish players play the SLM. Each player *i* does not know the true value of c, and has belief b_i about *c,* which is a degenerated distribution. To consider selfish player's incentive to choose *C*, Let c^* be the largest *c* satisfying $n\alpha w > w + c\alpha w$, that is, the largest integer equal to or less than $n - 1/\alpha$. Note that c^* is the threshold of selfish player's belief to choose *C*. To ensure weak dominance, in addition to $1/n < \alpha < 1$, we assume $\alpha \neq 1/(n\text{-}1), 1/(n\text{-}2), \ldots, 1/2$. Then, we obtain the following failure of cooperation due to incomplete information on type.

Proposition 2. *Suppose n players consisted of* $c \leq c^*$, *conditionally cooperative players and (n-c) selfish players and* $\alpha \neq 1/(n\text{-}1), 1/(n\text{-}2), \ldots, 1/2$. *Unanimous cooperation is achieved in one-shot under the SLM if and only if* $b_i \leq c^*$ *for every selfish player i.*

Proof. See Appendix.

A more attractive and paradoxical result will be obtained when we embed a structure between beliefs and the true number of conditional cooperators. Note that Proposition 2 says nothing about where beliefs come from. It seems plausible to assume

that beliefs correctly capture c at the aggregate level, even if realized beliefs are dispersed.

We assume that B-i) given c, for each selfish player i, b_i, is identically and independently drawn from a discrete uniform distribution between $c-d$ and $c+d$ and B-ii) the largest (smallest) possible belief is not above n (below 0): $c^*+d \le n$ and $c_{\min}-d \ge 0$, where $c_{\min} \ge 1$. B-ii) is to avoid some boundary problem. Then, a paradoxical result holds:

Proposition 3. *Assume* $n \ge 3$, $\alpha \ne 1/(n\text{-}1), 1/(n\text{-}2), \ldots, 1/2$, $c_{\min} \le c \le c^*$, *B-i) and B-ii). Then, unanimous cooperation is less likely to occur as the number of conditionally cooperative players increase.*

Proof. See Appendix. ■

The intuition behind the contrast between the homogeneous model (Proposition 1) and heterogeneous model (Propositions 2 and 3) is as follows. In the former model, just one *D* triggers the collapse all public good benefits since other players surely choose *Leave*, which makes *D* unprofitable. Hence, unanimous cooperation is maintained. In the latter model, however, as the number of conditionally cooperative players increase, the influence of selfish players on the true number also increases. Hence, selfish players are more likely to be tempted to *D*, since more conditionally cooperative players still *Stay* and the larger benefits remain after their choosing *D*.

Propositions 2 and 3 show the pessimistic view that social preference may hinder cooperation. Fortunately, if we assume repeated interaction, unanimous cooperation will be achieved with only one period of delay because regardless of their initial beliefs, players distinguish all other players' behavioral types based on the observed behavior in period 1.

Example 2. To see this, we suppose there is period 2 after Example 1. Now, consider belief updating before playing period 2. By observing player 3's *Stay*, players 1 and 2 update their beliefs to 1 since they know that "only player 3 is conditionally cooperative." On the other hand, player 3 updates his belief to 0 since player 1 chose *D* and player 2 chose *Leave*. To sum up, the updated beliefs are (1,1,0). Note that every player has the correct beliefs.

Given belief updating, consider period 2. Since player 3 realized that only she is conditionally cooperative, player 3 will choose *Leave* in both second-stage subgames. Hence, player 1's payoff will be 10<21, which suggests that player 1 cannot exploit player

3. Since *C* weakly dominates *D* in Table 6, player 1 chooses *C*. By a symmetric argument, *C* weakly dominates *D* for player 2. By definition, conditionally cooperative player 3 chooses *C*. Therefore, all players cooperate in period 2. From period 3 onward, beliefs are no longer updated. Hence, the same argument holds as that for period 2. Therefore, *CCC* is achieved from period 3 onward.

Examples 1 and 2 show the scenario where there are at least two *C* players. The second scenario is where there is only one *C* player, *i* (c=1). Note that player *i* chooses *Leave* regardless of his behavioral type. In this case, players other than *i* cannot distinguish which type *i* is. However, since players other than *i* chose *D*, all (*n*-1) players are revealed to be selfish. Hence, every player knows $c \le 1$. Under the assumption of $1 \le c^*$, all selfish players choose *C* from period 2 onward. The third scenario is unanimous cooperation achieved at the beginning.

Proposition 4. *If n players repeatedly play the SLM, for any* $\alpha, \alpha \neq 1/(n\text{-}1), 1/(n\text{-}2), ..., 1/2$, *any c,* $c_{\min} \le c \le c^*$, *any beliefs* $b_i, i=1,2,...,n-c$ *unanimous cooperation occurs in period 2 onward.*

Proof. See Appendix.■

6. Concluding remarks

We introduced the SLM for *n*-player prisoner's 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. In our SLM 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 such as Varian's (1994) compensation mechanism experiment for two-player prisoner's dilemma games (Andreoni and Varian, 1999; Charness et al., 2007), where the cooperation rate reached around 70% after dozens of repetitions. One limitation of the current study is that we do not compare the SLM directly with the other mechanisms in the literature such as the compensation mechanism. We can expect, however, that the former would yield a higher cooperation rate than the latter based on Saijo et al. (2016), since the authors provided experimental evidence that their approval mechanism, which is the origin of the SLM, outperforms the compensation mechanism in two-player prisoner's dilemma games. A companion paper Saijo and Masuda (2014) extends the SLM so that it works in *n*-player linear public good environment, considering

heterogeneity in value of the public good. Although the SLM can be regarded as a variant of Gerber et al.'s (2013) unanimous voting session IF4, our novelty lies in showing evidence and rebuilding model of mixed behavioral type as follows.

To explain the observed upward trend of cooperation rates in early periods, we conducted additional sessions with the strategy method and scrutinized individual choice data. We successfully verified the coexistence of selfish (50.0%) and conditionally cooperative players (14.7%), consistent with the accumulated evidence on conditionally cooperative subjects (Chaudhuri, 2011). In addition, we confirmed that the observed behavioral type distribution is not significantly different between elicitation methods.

Given the behavioral heterogeneity, we developed a novel model incorporating two behavioral types and heterogeneous beliefs, giving rich predictions successfully. First, we show that increasing cooperativeness among players may paradoxically hinder efficiency. This message is in contrast to Kosfeld et al. (2009), rather in line with Kube et al. (2015).

Second, we show that players can achieve unanimous coopearaion through learning others' types in repeated interaction. Note that the latter prediction is consistent with our data. Moreover, our model has distinctive features from the one in Kube et al. (2015) based on Fehr and Schmidt (1999) in the sense that we do not need hidden preference parameters on equity, and the uncertainty of our model is summarized into type population. We should emphasize that our efficiency-reducing result comes solely from behavioral heterogeneity, while the argument in Kube et al. (2015) rely on tension between efficiency and equit due to heterogeneity of marginal benefit of the public good. Indeed, the implications of our experiment shed light on the importance of incorporating behavioral heterogeneity into mechanism design, in line with Andreoni and Varian (1999), Charness et al. (2007), and Levati and Neugebauer (2004). One fruitful way would be considering mechanisms for continuum of conditionally cooperative players proposed in Andreoni and Samuelson (2006) where both selfish and altruistic players are extreme cases.

Appendix. Proofs

Proof of Proposition 1.

Let $n \geq 2$ and $\alpha \in (1/n, 1)$. Assume that all players are selfish. 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 chose *C*. Then, by construction of the SLM, second-stage mover chooses *Leave* because it gains $(1-\alpha)w$ than *Stay*. Then, this player gets 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 get αnw if all of *n* players

choose *C* in the first stage, and will get 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.

Note that a selfish player is indifferent between *C* then *Leave* and *D*, unless she can be a pivotal, that is, the player who induces conditional cooperators *Stay* by choosing *C*. By definition of l) and m) of conditionally cooperative player, that happens only when there is only one conditionally cooperative player and all of other (*n*-2) selfish players choose *D*. Denote this by $\cdot CD...D$, where the dot indicates either *C* or *D* of selfish player of our interest. Being a pivotal player is profitable, since given $\cdot CD...D$, choosing *C* then *Leave* yields $w + \alpha w$ while choosing *D* yields w.

Take any $\alpha, \alpha \neq 1/(n\text{-}1), 1/(n\text{-}2), ..., 1/2$ and any *c*. Take any selfish player $i = 1, 2, \ldots, n - c$. Suppose first $b_i > c^*$. By $c^* \geq 1$, we have $b_i \geq 2$. When all players except for *i* choose *C*, by definition of c^*, $w + c^* \alpha w < n\alpha w < w + b_i \alpha w$. That is, choosing *D* is better. When not all of players except for *i* choose *C*, by $b_i \geq 2$, player *i* believes that she cannot be a pivotal. Hence, player *i* is indifferent between *C* then *Leave* and *D*. Therefore, *D* weakly dominates *C*.

Suppose next that $b_i \leq c^*$. When all players except for *i* choose *C*, by definition of c^*, $n\alpha w > w + c^* \alpha w \geq w + b_i \alpha w$. That is, choosing *C* is better. When not all of players except for *i* choose *C*, there are two cases. If $b_i = 1$, player *i* is indifferent between *C* then *Leave* and *D* except at $\cdot CD...D$ by pivotal argument. If $b_i \neq 1$, player *i* believes that she cannot be a pivotal. Hence, player *i* is indifferent between *C* then *Leave* and *D* for any constellation of *C* and *D* where the number of C players is between b_i and *n*-1. Therefore, in both case of b_i, *C* weakly dominates *D*. ∎

Proof of Proposition 3

Proof. It suffices to show that the probability that all (*n*-*c*) selfish players choose *C*, denoted by *P*(c), is decreasing in *c*. For analytical convenience, we regard *c* as a continuous variable. Take any $n \geq 3$, any $\alpha \neq 1/(n\text{-}1), 1/(n\text{-}2), ..., 1/2,$ and any $c_{\min} \leq c \leq c^*$. Assume B-i) and B-ii) hold. Note that selfish player *i* chooses *C* if and only if $b_i \leq c^*$. Then, $P(c) = \Pr(b \leq c^* \text{ for all selfish players} \mid c)$. Consider the case $c + d < c^*$. Then, $b_i \leq c + d < c^*$ for all selfish *i*. Hence, $P(c) = 1$. Consider the case $c + d \geq c^*$.

Then, $P(c) = \Pr(b_1 \leq c^* \text{ for player } 1 \mid c)^{n-c} = x^y$. where $x = \dfrac{c^* - (c - d)}{2d}$ and $y = n - c$.

Hence, $\dfrac{dP}{dc} = \dfrac{\partial P}{\partial x}\dfrac{dx}{dc} + \dfrac{\partial P}{\partial y}\dfrac{dy}{dc} = yx^{y-1}\left(\dfrac{-1}{2d}\right) + x^y \log x(-1) = (-x^{y-1})Q(x, y),$

where $Q(x,y)=\frac{y}{2d}+x\log x$. $\frac{dP}{dc}<0 \Leftrightarrow Q(x,y)>0$. We show $Q(x,y)>0$. First, by $n \geq (c^*+d)$, $\frac{y}{2d}=\frac{n-c}{2d}+\frac{c^*+d}{2d}-\frac{c^*+d}{2d}=\frac{c^*-(c-d)}{2d}+\frac{n-(c^*+d)}{2d} \geq x+0=x$. Second, by $c^* \geq c$, $x=\frac{d+(c^*-c)}{2d} \geq \frac{1}{2}$. Third, $\frac{d\{x(1+\log x)\}}{dx}=2+\log x>0$ for $x \geq \frac{1}{2}$. Hence, $Q(x,y) \geq x(1+\log x) \geq \frac{1}{2}(1+\log\frac{1}{2})>0$. Therefore, *P* is monotonically decreasing in *c*. ■

Proof of Proposition 4.

Take any $\alpha, \alpha \neq 1/(n\text{-}1), 1/(n\text{-}2), \ldots, 1/2,$ and any $c, 1 \leq c \leq c^*$. Note that $c^* < n-1/\alpha$. There are three possible cases.

Case 1. $b_i \leq c^*$ for all selfish players i. In this case, unanimous cooperation is achieved for every period. No one updates one's beliefs because no one observes the second-stage choices.

Case 2. $b_i > c^*$ for all selfish players i. In this case, all conditionally cooperative players choose *C*, while all selfish players choose *D*. If there are at least two conditionally cooperative players, they perfectly reveal their type by choosing *Stay*. If there is only one conditionally cooperative player, her *Leave* does not reveal her type, since in that situation both types choose *Leave*. In both cases, all players know $c \leq 1$. From $c^* \geq 1$, unanimous cooperation is achieved from period 2 onward.

Case 3. $b_j \leq c^* < b_i$. for a pair of players *i* and *j*. Since *j* chooses *C*, conditionally cooperative players choose *Stay*, without distinguishing whether *C* is chosen by conditionally cooperative or selfish players. Hence, all players perfectly reveal their types in period 1. Since beliefs are correctly updated, unanimous cooperation is achieved from the next period. ■

Acknowledgments

We thank Takako Greve, Chiaki Hara, Tadashi Sekiguchi, Kazumi Shimizu. Keiko Takaoka also provided outstanding research assistance. We also thank the participants of the game theory seminar at the Institute of Economic Research, Kyoto University, Asia Economic Institutes Workshop, the 25th Annual Meeting of the Japanese Society for Mathematical Biology. 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.

Masuda is grateful for the financial support from K-CONNEX and the Japan Society for the Promotion of Science for the postdoctoral fellowship.

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.

Bearman, J. (2007). More giving together: The Growth and impact of giving circles and shared giving. Forum of Regional Associations of Grantmakers.

Blanco, M., Engelmann, D., Koch, A.K., Normann, H.T., 2014. Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. Games and Economic Behavior, 87, 122-135.

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

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.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117 (3), 817–869.

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.

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.

Eikenberry, A. M., Bearman, J., Han, H., Brown, M., & Jensen, C. (2009). The impact of giving together: Giving circles' influence on members' philanthropic and civic

behaviors, knowledge and attitudes.

Fehr, E., Schmidt, K., 1999. Fairness and retaliation: the economics of reciprocity. Quarterly Journal of Economics 114, 817–868.

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.

Fréchette, G. R., Kagel, J. H., & Morelli, M. (2012). Pork versus public goods: an experimental study of public good provision within a legislative bargaining framework. Economic Theory, 49(3), 779-800.

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.

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

Kesternich, M., Lange, A., Sturm, B. 2014. The impact of burden sharing rules on the voluntary provision of public goods. Journal of Economic Behavior & Organization 105 (September), 107-123.

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.

Lindquist, S., 2007. Ex Interim Voting: An Experimental Study of Referendums for Public-Good Provision: Comment. Journal of Institutional and Theoretical Economics / Zeitschrift Für Die Gesamte Staatswissenschaft, 163(1), 81-83.

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.

Saijo, T., Masuda, T., 2014. The simplest solution to the free-rider problem: Theory and experiment. Unpublished manuscript.

Saijo, T., Okano, Y., Yamakawa, T., 2016. The mate choice mechanism experiment: A solution to prisoner's dilemma. KUT-SDE working paper series no. 2015-12 (reviesed), 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.