Approval Mechanism to Solve Prisoner's Dilemma: Comparison with Varian's Compensation Mechanism

Tatsuyoshi Saijo[1,2,3], Takehito Masuda[4,5] and Takafumi Yamakawa

October 2017

Abstract

After having played a prisoner's dilemma, players can approve or reject the other's choice of cooperation or defection. If both players approve the other's choice, the outcome is just the result of the chosen strategies in the prisoner's dilemma; however, if either rejects the other's choice, the outcome is the result of mutual defection in the prisoner's dilemma. In theory, such an approval mechanism implements cooperation in backward elimination of weakly dominated strategies, although this is not the case in subgame perfect Nash equilibrium. By contrast, Varian's (1994; A solution to the problem of externalities when agents are well-informed. Am Econ Rev 84(5): 1278-1293) compensation mechanism implements cooperation in the latter but not in the former, which motivates the present study. The approval mechanism sessions yield a cooperation rate of 90% in the first period and 93.2% across periods, while the compensation mechanism sessions yield a rate of 63.3% in the first period and 75.2% across periods, indicating a significant difference. In addition, the backward elimination of weakly dominated strategies better predicts subjects' behavior than subgame perfect Nash equilibrium in both mechanism sessions.

JEL codes: C72, C73, C92, D74, P43

Keywords: prisoner's dilemma, approval mechanism, cooperation, backward elimination, weakly dominated strategies, laboratory experiment, Selten's index

1. Introduction

Aligning participants' individual interests with the collective ones is key to designing mechanisms to overcome a prisoner's dilemma (PD) and achieve efficient public goods provision in general. Despite the long history of theoretical approaches in institution design, "behavioral

[1] Corresponding Author. E-mail: tatsuyoshisaijo@gmail.com Phone: +81-88-821-7155 Fax: +81-88-821-7198

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

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

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

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

assumption made in theory is most seriously challenged" (Chen 2008, p. 626) in laboratory experiments. Hence, an effective way to make progress in institution design is to determine what combination of mechanisms and behavioral theory works in an experimental setting that is as simple as possible.

Subgame perfect Nash equilibrium (SPNE) is an example of a behavioral assumption that has been challenged in social dilemma experiments. Varian (1994) proposed a mechanism to attain efficiency in SPNE, calling it a compensation mechanism (CM). The mechanism gives players an opportunity to offer transfers contingent on cooperative action prior to playing the underlying game. Applying the CM to the prisoner's dilemma, one calculates that, at equilibrium, players mutually cooperate after offering to compensate each other the gain from unilateral defection.[6] Despite the theoretical properties of the CM, experimental evidence appears to confront theory. In the PD with Andreoni and Varian's (1999) CM experiment, about one-third of the subjects faced difficulties in achieving SPNE even after repeating the play 20 times. More recent studies have reported substantial deviations from SPNE in CM experiments with more complex externality settings.[7] Thus, designing multi-stage mechanisms that work well in the laboratory remain a challenging problem.

In this study, we attempt to address this challenge by introducing the *approval mechanism* (AM). Consider adding an approval stage after the PD, where each subject can approve ("*yes*") or disapprove ("*no*") the other's choice of strategy in the first stage: if both approve the other's strategy, the outcome is just the result of the chosen strategies in the PD; however, if either disapproves, the outcome is the result of mutual defection in the PD.[8] The *AM* implements cooperation in *backward elimination of weakly dominated strategies* (BEWDS) but not in SPNE (Properties 1 and 2). By contrast, the CM implements cooperation in SPNE but not in BEWDS (Property 3). These contrasting features of the AM and CM motivate us to compare them experimentally.

The experimental data suggest that the AM works better than the CM. We employed a between-subject and complete-stranger design. The AM sessions yielded a cooperation rate of 90% in the first period and 93.2% across periods, whereas the CM sessions yielded a rate of 63.3% in the first period and 75.2% across periods. We find a significant difference in the mean cooperation rate between the AM and CM.

A classification of group-level data reveals that BEWDS is a better predictor of subjects'

[6] This holds true provided the payoff asymmetry in the underlying prisoner's dilemma is not so large (Qin 2005).

[7] See Hamaguchi et al. Saijo (2003), Bracht et al. (2008), and Midler et al. (2015). For Moore-Repullo SPNE mechanism experiments, see Fehr et al. (2014).

[8] Selten (1975) and Kalai (1981) applied BEWDS to game theory.

behavior in both the AM and CM. Under the AM, the BEWDS path is unique (Property 1), but SPNE paths are multiple (Property 2). Moreover, SPNE paths include a BEWDS path. However, the opposite holds true for the CM (Property 3). These properties motivate us to use Selten's index of predictive success (Selten 1991), which captures both the correctness and parsimony of equilibrium predictions. The data for the AM show that Selten's index for BEWDS is about twice that for SPNE, suggesting that BEWDS is a better predictor. The data for the CM, on the other hand, show that, although BEWDS yields a slightly higher Selten's index than SPNE, neither equilibrium concept fits the data well.

The contributions of this paper to the mechanism experiment literature are threefold. First, we successfully designed a two-stage mechanism that works better than the CM in laboratory experiments. The result could be attributed to the simplicity of backward thinking under the AM. That is, the AM subjects have to check only four second stages, while the CM subjects have a large number of transfer options and hence face a heavy cognitive burden to think backwardly, which hinders efficiency, as reported by Midler et al. (2015). Moreover, a noteworthy feature of the AM is that it does not utilize private punishment and/or reward technologies, which are sometimes assumed in the literature (Fehr and Gächter 2000; Varian 1994). However, since personal punishment (or bribe) is generally prohibited in modern societies or legal systems, we view this as a strength of the AM.[9] On the other hand, under the AM, when either player disapproves the choice of the other, the public good will not be provided, and thus, the money is simply returned to the contributor. Second, our experimental design with symmetric PD allows us to detect genuine difficulties subjects faced in finding SPNE under the CM. Although Charness et al. (2007) point out that equity concerns regarding final payoffs hinder the achievement of SPNE in their CM experiment with asymmetric PDs, this is not the case under a symmetric PD. Third, our analysis on group-level data sheds light on an unexplored topic: the *affinity* between the mechanism and behavior. Since subjects behave almost consistently with BEWDS under the AM but not the CM, the possibility that mechanisms cause subjects to demonstrate a particular behavior is a plausible consideration.

The remainder of this paper is organized as follows. Section 2 explains the AM, and section 3 identifies BEWDS and SPNE under the AM. Section 4 introduces the CM. Section 5 presents the experimental design. Section 6 discusses the experimental results. Section 7 concludes the paper.

[9] In this vein, Saijo, Okano, and Yamakawa's (2016) working paper characterizes the AM using certain no-punishment and no-coercion conditions. Kimbrough and Sheremeta (2013) present situations in which side payments assumed in the CM might not be legal: collusion in a market, patent races, and R&D competition. Guala (2012) points out that there is little anthropological evidence that humankind has used private punishment.

2. Approval Mechanism

The AM consists of two stages. In the first stage, players 1 and 2 face a typical PD game such as that presented in Fig. 1. Both players must choose either cooperation (*C*) or defection (*D*). While there might be many ways to interpret the matrix in Fig. 1, a typical interpretation in public economics is the payoff matrix of the voluntary contribution mechanism for the provision of a public good. Each player has $10 (or initial endowment *w*) initially, and must decide whether to contribute the whole $10 (cooperate) or nothing (defect). The sum of the contribution is multiplied by $\alpha \in (0.5, 1)$, which is 0.7 in Fig. 1, and the benefit is derived by both players, which indicates the non-rivalry over the public good. If both contribute, then the benefit to each player is (10 + 10) × 0.7 = 14. If either contributes, the contributor's benefit is 10 × 0.7 = 7, while the non-contributor's benefit, including the $10 remaining, is 10 + 7 = 17. Therefore, the payoff matrix in Fig. 1 maintains a linear structure, with non-contribution (*D*) as the dominant strategy. The bold numbers in the lower right cell denote the equilibrium payoff.

		Player 2	
		C	*D*
Player 1	*C*	14,14	7,17
	D	17,7	**10,10**

Fig. 1 PD game

Consider now the following second stage (approval): if both approve the other choice in the first stage, then the payoff (or outcome) is what they choose in the PD stage. Otherwise, the payoff is (10,10), which corresponds to (*D,D*) in the first stage.[10]

There are many examples of the AM. Consider a merger or joint project undertaken between two companies. They must cooperatively propose plans in the first stage, with each then facing an approval decision in the second. Another example is the two-party political system. Each party chooses either cooperation (or compromise) or defection (or insistence on one's own policy), and the parliamentary body then approves or disapproves the choice. The bicameral system also

[10] The AM also differs from the money back guarantee mechanism as follows: If either player, but not both, chooses *C*, then the $10 contribution is returned to the cooperator. The money back guarantee mechanism cannot generate (7,17) when (*C,D*) occurs in the first stage of the AM, while both choose *y* in the second stage of the AM. The same argument applies to Brams and Kilgour's (2009) idea to vote between the outcomes of all-*C* and all-*D* in order to overcome PD. The AM has an advantage in terms of welfare compared to the money back guarantee mechanism and Brams and Kilgour's (2009) voting. To see this, consider a case where only player 1 is so altruistic that player 1's utility is given by (player 1's material payoff)+ ρ (player 2's material payoff), $\rho > 4/3$. Assume that players know each other's payoff function. Then, the outcome of *CD* maximizes the sum of utilities of two players among four possible outcomes of the PD in Figure 1. The AM implements *CD,* while neither the money back guarantee mechanism nor Brams and Kilgour's (2009) voting does.

has two stages. For example, in the negotiation process at the United Nations, negotiators from relevant countries first assemble to determine compromises, and high-ranking officials, such as presidents and prime ministers, then approve or disapprove these decisions. These examples demonstrate that adding an approval stage to resolve conflicts is widely used in society.

3. Predictions under the AM

3.1. Backward Elimination of Weakly Dominated Strategies

Backward elimination of weakly dominated strategies (BEWDS), which was also adopted by, for example, Kalai (1981), requires two properties: (i) subgame perfection and (ii) an understanding that players do not choose weakly dominated strategies in each subgame or in the reduced normal form game. We now explore the subgame starting from *CC* in Fig. 2.

Player 2

		Y	*n*
Player 1	*y*	***14,14***	10,10
	n	10,10	***10,10***

Subgame *CC*

		y	*n*
Player 1	*y*	7,17	10,10
	n	***10,10***	***10,10***

Subgame *CD*

		y	*n*
Player 1	*y*	17,7	***10,10***
	n	10,10	***10,10***

Subgame *DC*

		y	*n*
Player 1	*y*	***10,10***	***10,10***
	n	***10,10***	***10,10***

Subgame *DD*

Fig. 2 Four subgames in the AM

Player 2

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

(i) *yy* in subgame *CC*

Player 2

		C	*D*
Player 1	*C*	***10,10***	***10,10***
	D	***10,10***	***10,10***

(ii) *nn* in subgame *CC*

Fig. 3 Reduced normal form games

Suppose that player 1 chooses *y*. Then, the payoff of player 1 is 14 when player 2 chooses *y* and 10 when player 2 chooses *n*. Thus, the vector of possible payoffs is (14,10). The vector (10,10) then corresponds to player 1's choice of *n.* We say choice *x* associated with a possible payoff vector (*u*,*v*) *weakly dominates* choice *z* associated with a possible payoff vector (*s*,*t*) if $u \geq s$ and $v \geq t$ with at least one strict inequality. Since *y* weakly dominates *n*, *n* should not be chosen. Therefore, *yy* is realized in subgame *CC*. The cell with the bold black italic values denotes this outcome. Similarly, *ny* in subgame *CD* and *yn* in subgame *DC* are realized. In subgame *DD*, since no weakly dominated strategy exists, *yy*, *yn*, *ny*, or *nn* is realized. Given the realized strategies in all subgames, we have the reduced normal form game (Fig. 3-(i)). In this game, *C* weakly dominates

D for both players, and hence, *CC* is the realized outcome. The same is true as long as the marginal benefit of cooperation is between 0.5 and 1.

Property 1. The AM implements cooperation in BEWDS.

3.2. *Subgame Perfect Nash Equilibrium*

Now, consider the SPNEs in the AM. To do this, we examine each subgame in the second stage shown in Fig. 2. The NEs in subgame *CC* are *yy* and *nn* because of the indifference among *yn*, *ny*, and *nn*. Furthermore, *ny* and *nn* in subgame *CD*, *yn*, and *nn* in subgame *DC*, and all four combinations in subgame *DD* are NEs. The cells with the bold gray italics indicate the NEs that are not BEWDS outcomes. Since the NEs in subgame *CC* are *yy* and *nn*, there are two reduced normal form games (Fig. 3). (*C,C*) and (*D,D*) are NEs in Fig. 3-(i), and all combinations are NEs in Fig. 3-(ii). For example, *CDny* is an SPNE path.[11,12] Thus, we have the following property.

Property 2. The AM cannot implement cooperation in SPNE.

Although we can argue other equilibrium selections, in what follows, we restrict our attention to BEWDS and SPNE given the main purpose of this study.[13]

4. Compensation Mechanism

Next, we explain the *compensation mechanism* (CM) developed by Varian (1994). In the first stage of the CM, each subject could offer to pay the other subject *before* the second PD stage if the latter cooperates Then, both cooperate in the unique SPNE, as long as the payoff asymmetry in PD is within a certain range (Qin 2005). In Andreoni and Varian (1999), random-matched groups of two play a slightly asymmetric PD game for the first 15 periods and, then, play the CM from the 16th to the 40th periods. The cooperation rate of the CM was 50.5%, which is contrary to the SPNE prediction.

In the CM using PD in Fig. 1, the equilibrium offers are either three or four in the CM, and

[11] A *path* is (player 1's choice between *C* and *D*, player 2's choice between *C* and *D*, player 1's choice between *y* and *n*, player 2's choice between *y* and *n*). For simplicity, hereafter, we write a path as *CDny* instead of (*C,D,n,y*).

[12] In the AM, we have 96 SPNEs. See Saijo et al. (2016) for more details.

[13] The risk dominance criterion, even with subgame perfection, cannot select the strategy uniquely, unlike in the case of BEWDS. The off-path prediction by risk dominance is redundant. To observe this, consider subgame *CD* in Fig. 2. According to risk dominance, a NE with greater product deviation loss is more likely to occur (e.g., Blonski and Spagnolo 2015). Since the product of loss deviating from NE (*n,y*) and that of (*n,n*) are zero, both are predicted in the light of risk dominance. However, seven out of the eight observed choices in subgame *CD* are (*n,y*), as reported in Table 2.

then both choose cooperation in the PD stage. On the other hand, all possible combinations, (*C*,*C*), (*C*,*D*), (*D*,*C*), and (*D*,*D*), fall under BEWDS with an equilibrium offer of three. The BEWDS outcomes are 33*CC*, 33*CD*, 33*DC*, 33*DD*, 34*CC*, 34*CD*, 43*CC*, 43*DC*, and 44*CC*.[14] In sum, we have the following property.

Property 3. The CM implements cooperation in SPNE, but not in BEWDS.

5. Experimental Procedures

We conducted the experiments at Osaka University in November 2009, March 2010, October and November 2011, and January 2012. The AM and CM had three sessions each, and PD had one session. In each session, 20 subjects played the game in 19 periods. We created 10 pairs out of the 20 subjects and seated them at computer terminals in each session. We used the z-Tree program (Fischbacher 2007). We employed complete stranger matching.[15] No subjects participated in more than one session. We recruited these subjects through whole-campus advertisements. Subjects were told that there would be an opportunity to earn money in a research experiment. Communication among the subjects was prohibited. Each subject received an instruction sheet and record sheet. The instruction was read aloud by the same experimenter.

We now explain the AM. Before the payment periods began, we allowed the subjects five minutes to examine the payoff table and consider their strategies. When the period started, each subject selected *A* (defection) or *B* (cooperation) in the choice (or PD) stage, and then fed their choice into a computer and made a mention on the record sheet. Then, the subjects wrote the reason for their choice in a small box on the record sheet. Next was the decision (or approval) stage. Knowing the other's choice, each subject chose to either "accept" or "reject" the other's choice, and then reported the decision on the computer and record sheet. Then, each subject wrote the reason in a small box. Once all subjects finished the task, each of them could see "your decision," "the other's decision," "your choice," "the other's choice," "your points," and "the other's points" on the computer screen. Subjects got no information on the choices and decisions made in the other groups. This ended one period. The experiment without the decision stage became the PD. After 19 periods of play, subjects filled in questionnaire sheets.

In the CM sessions, when a period begins, subjects would proceed to the transfer stage, where they choose how many points to transfer to the opponent if the opponent chooses *B* in the second (choice) stage. The transfer must be a nonnegative integer with an increment of 100. In the

[14] These results are due to the discreteness of strategies.

[15] The pairings were anonymous and determined in advance so that no two subjects were paired more than once.

choice stage, subjects choose *A* or *B* knowing the pair of offered transfers.

6. Experimental Results

6.1. Comparison of AM with CM

Fig. 4 illustrates the cooperation rates of the AM, CM, and PDs per period. We use the *ex post* cooperation rate in the AM experiments.

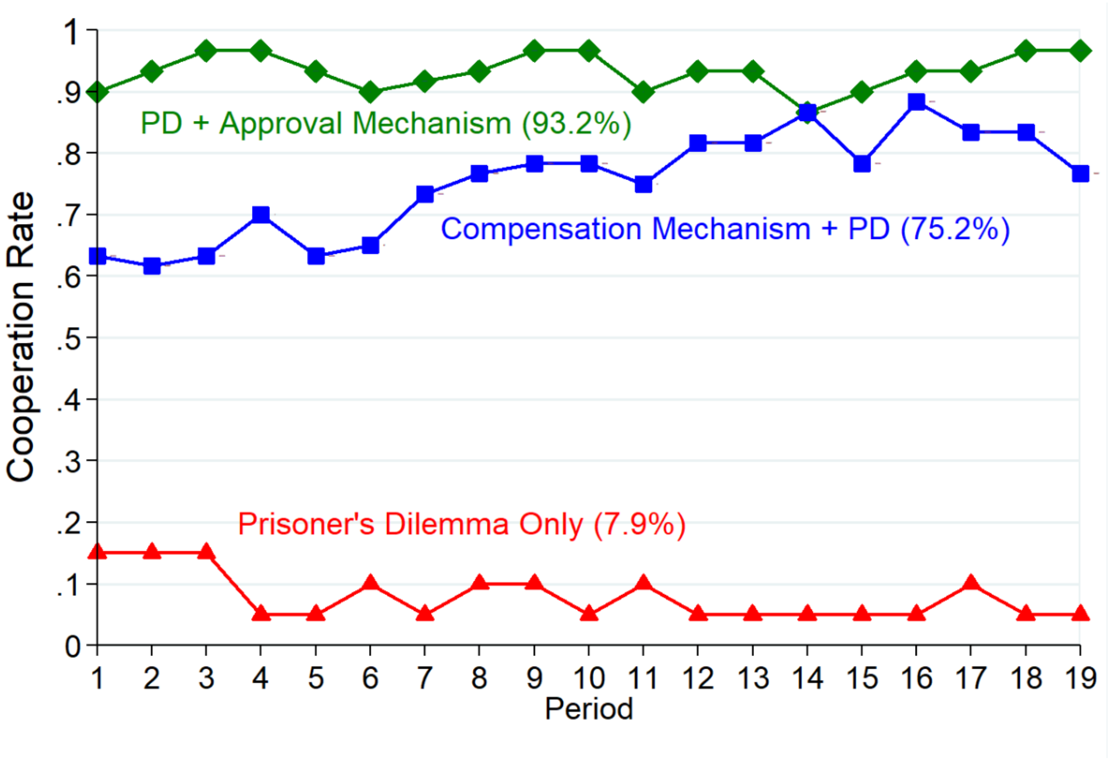

Fig. 4 Cooperation rates by treatment

For example, if both chose *C* in the choice stage and one of the subjects disapproved the other's choice in the decision stage, then we did not count their choices as cooperation. The AM achieves a high cooperation rate for periods 1–19; the cooperation rate is 90% in period 1, and at least 90% in all periods except period 14, with an average rate of 93.2%. The AM yielded a significantly higher average cooperation rate than the CM (p-value < 0.001, Wilcoxon rank-sum test).[16] The latter yielded 63.3% in the first period and 75.2% across periods. The period-by-period Chi-square test results in Table 1 indicate that the gap in the cooperation rate between the AM

[16] We used Andreoni and Miller's (1993) method. We first calculated the average cooperation rate for each subject across periods, followed by the test statistic using the averages to eliminate cross-period correlation.

and CM was sustained, especially in the first 10 periods.

As a control treatment, the PD obtained a 7.9% cooperation rate; specifically, it is 11% for the first five periods, but declines to 6% in the last five. Further, no *CC* is observed among the 190 pairs of choices.[17] Both the AM and CM promote cooperation compared to the PD (p-value < 0.001 for both AM vs. PD and CM vs. PD, Wilcoxon rank-sum test).[18]

Period	1	2	3	4	5	6	7	8	9	10
AM vs. CM	0.001	0.000	0.000	0.000	0.000	0.001	0.008	0.011	0.002	0.002

Period	11	12	13	14	15	16	17	18	19
AM vs. CM	0.031	0.053	0.053	1.000	0.080	0.343	0.088	0.015	0.001

Table 1. p-values of chi-square test for each period

6.2. *BEWDS predicts data more successfully than SPNE*

We determine whether BEWDS or SPNE is a better predictor of subjects' behavior by analyzing individual choices. We use a path as a unit of prediction and observation. Note that the above-mentioned theoretical properties make it difficult to establish this for the following reasons. Consider the AM. While the BEWDS path is unique (Property 1), SPNE paths are multiple (Property 2). Moreover, SPNE paths include a BEWDS path. However, the opposite holds true for the CM (Property 3). With such multiplicity and overlapping equilibrium sets, it is important to consider both where the data accumulate and the degree to which the predictions are sharp.

Selten's index of predictive success is a measure introduced and axiomatized by Selten (1991) to balance both concerns.[19] Let a mechanism and an equilibrium concept (BEWDS or SPNE) be given. The index is the difference between two components that respectively correspond to the descriptive power and parsimony of the equilibrium prediction. The first component is the pass rate r, which is the proportion of correctly predicted observations. We give an example of r using Table 2, which summarizes predictions, observations, and Selten's indices. Consider the AM sessions shown in panel (a). Note that there are 19 (subjects) x 10 (groups) x 3 (sessions) = 570 observations. The second column shows that the BEWDS prediction is only *CCyy* (Property 1). Since 531 out of 570 observations are *CCyy* paths, we have r = 531/570 = 0.932 for BEWDS. The second component of Selten's index is area a, which is the ratio of the number of equilibrium

[17] The cooperation rate in the PD is slightly lower than that recorded in previous experiments. For example, Cooper et al. (1996) and Andreoni and Miller (1993) found 20% and 18% cooperation rates, respectively.

[18] Additional experiments on the AM and PD where subjects play the game in only one period provide similar results, suggesting that the AM works without repetition. The data and results are available upon request.

[19] Selten's index has been used in the literature on testing revealed preference. Our notations follow Beatty and Crawford (2011). Forsythe et al. (1996) applied Selten's index to a voting game experiment.

predictions to the total number of possible paths. Note that for each player, there are two choices in each stage (*C* or *D*, *y* or *n*) and there are $(2 \cdot 2)^2 = 16$ possible paths. Since the BEWDS prediction is only *CCyy*, we have *a* = 1/16 = 0.063 for BEWDS. Hence, under the AM, Selten's index for BEWDS *s* = *r* – *a* = 0.932 – 0.063 = 0.869. Note that a larger *s* is better and –1 < *s* < 1.

Path	BEWDS	SPNE	Obs.	Pass rate *r*	Area *a*	Selten's Index *s* = *r* - *a*
CCyy	Yes	Yes	531	0.932 (=531/570))	0.063 (=1/16)	0.869
CCnn	No	Yes	0	0.990 (=564/570)	0.625 (=10/16)	0.365
*CDny**	No	Yes	28			
*CDnn**	No	Yes	4			
DDyy	No	Yes	1			
*DDyn**	No	Yes	0			
DDnn	No	Yes	0			

(a) AM

Path	BEWDS	SPNE	Obs.	Pass rate *r*	Area *a*	Selten's Index *s* = *r* - *a*
33*CC*	Yes	Yes	48	0.423 (=241/570)	0.04 (=4/100)	0.383
34*CC**	Yes	Yes	58			
44*CC*	Yes	Yes	135			
33*CD**	Yes	No	5	0.577 (=329/570)	0.09 (=9/100)	0.487
33*DD*	Yes	No	0			
34*CD**	Yes	No	83			

(b) CM

Notes. a) We count every asymmetric path (*) twice to calculate area *a* because of its permutation. b) In the CM treatment, we compute area *a* by assuming that each player chooses a transfer from {0,100,200,300,400}. The result holds for any number of possible actions in the first stage more than five.

Table 2. BEWDS vs. SPNE by mechanism

Consider the AM and SPNE. The mark "*"indicates that the path is asymmetric, and thus, we must count its permutation. There are 10 SPNE paths (*CCyy, CCnn, CDny, DCyn, CDnn, DCnn, DDyy, DDyn, DDny*, and *DDnn*), which explain 531 + 28 + 4 + 1 = 564 observed paths. Thus, *r* = 564⁄570 = 0.990. On the other hand, *a* = 10/16 = 0.625. Hence, Selten's index *s* = *r* – *a* = 0.990 – 0.625 = 0.365 for SPNE, which is much less than the index for BEWDS. Hence, we conclude that BEWDS is a better predictor than SPNE under the AM. Note that for SPNE, *r* is higher largely by the observations of *CDny*. It is natural to consider that player 2 tried to exploit player 1 in this path. Note that, on the other hand, in order *CDny* to be an SPNE path, *nn* must occur in subgame *CC*, which are dominated actions and never occurred.

Next, we examine the CM case shown in panel (b) of Table 2. Note that the second and third columns show that the SPNE paths are included in the BEWDS paths (Property 3). BEWDS paths explain 57.7% of the path data, while SPNE paths cover 42.3%. A frequently observed non-SPNE path is *34CD*, which survives under BEWDS since player 2 is indifferent between *C* and *D* in subgame *34*. However, *CC* must be chosen so that *34* constitutes the SPNE path. To compute area *a* in the CM, we must fix the number of possible paths and, thus, the number of first-stage alternatives. Table 2 (b) presents the area when each player can choose a transfer from {0,100,200,300,400}, while in the experiment, subjects can offer any nonnegative multiple of 100. Hence, the number of all possible paths under the CM is $(5\cdot 2)^2 = 100$. Then, the area for BEWDS is 9/100 = 0.09, while that for SPNE is 4/100 = 0.04. Here, too, BEWDS yields a higher Selten's index of 0.577 − 0.09 = 0.487 than SPNE, with an index of 0.423 − 0.04 = 0.383. However, it is noteworthy that neither SPNE nor BEWDS performs well in the CM.

7. Concluding Remarks

The present study theoretically showed that the AM implements cooperation in BEWDS. We found that the AM promotes cooperation significantly in a PD experiment, compared to the CM implementing cooperation in SPNE. Utilizing Selten's index to predictive success, we also demonstrated that BEWDS well explains data in the AM, but neither BEWDS nor SPNE does in the CM. Interestingly, this could be due to a large number of transfer options, which puts a heavy cognitive load on the subjects to find equilibria, as shown in Midler et al.'s (2015) CM experiment.

Undoubtedly, the AM does not always solve all PD games. First, participants must agree to use the mechanism as mechanism designers in all economics fields implicitly presume. Second, the mechanism might need monitoring devices and/or an enforcing power; otherwise, a participant might not perform the task described in *C* even after two participants choose *C* and *y*. Third, we cannot apply the mechanism if the contents of *C* have not yet been agreed upon. For example, although many researchers have used global warming as an example of PD, countries have been negotiating actions to address climate change for over 20 years under the United Nations Framework Convention on Climate Change without having agreed on what exactly *C* should be.[20]

[20] Although many people advocated the Paris Accord, the national pledges by countries to cut emissions are voluntary and do not involve penalty. "At best, scientists who have analyzed it say it will cut global greenhouse

We close with a brief mention of related research projects. Masuda et al. (2014) designed a minimum AM for a two-person voluntary contribution game for the provision of a public good, avoiding the failure of AM, as in the case of Banks et al.'s (1988) unanimous voting, and found that the minimum AM implements the Pareto-efficient outcome theoretically and experimentally. Second, for an *n*-player PD situation, Huang et al. (2016) designed the stay-leave mechanism, which leads to a behavioral mixture: BEWDS players and conditional cooperators. Third, consider a situation with at least three strategies and participants. Although a number of studies have explored this environment (Plott and Smith 2008), the gap between theory and experiments is yet to be bridged. Fourth, Saijo and Shen (2015) report that the AM works well even with an asymmetric PD, whereas the CM does not. Finally, exploring the *affinity* of mechanism and behavior to solve a specific problem is an important future research agenda, given that BEWDS explain data in the AM but not in the CM and subjects' equity concerns emerged through a transfer stage of the CM in Charness et al. (2007).

Acknowledgement

We thank the associate editor and two anonymous referees for their useful comments and suggestions that have improved the quality of the paper. The authors are very grateful to Yoshitaka Okano for his contributions to previous versions of this paper. Masuda is also grateful for helpful comments from Ryan Tierney. This research was supported by the Suntory Foundation; the Joint Usage/Research Center at ISER, Osaka University; Scientific Research A (24243028) and Challenging Exploratory Research (16K13354) of Kakenhi; and "Experimental Social Sciences: Toward Experimentally-based New Social Sciences for the 21st Century," a project funded by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science and Culture of Japan. We also thank the Economics Department, Osaka University, for allowing us to use their computer lab.

gas emissions by about half the required amount to avert a potential increase in atmospheric temperatures of 2 degrees Celsius or 3.6 degrees Fahrenheit"
(NYTimes, http://www.nytimes.com/2015/12/13/world/europe/climate-change-accord-paris.html).

References

Andreoni J, Varian H (1999) Preplay contracting in the prisoners' dilemma. Proc Natl Acad Sci 96(19): 10933-10938.

Andreoni J, Miller JH (1993) Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. Econ J 103(418): 570-585.

Beatty T, Crawford I (2011) How demanding is the revealed preference approach to demand? Am Econ Rev 101(6): 2782-2795.

Bracht J, Figuieres C, Ratto M (2008) Relative performance of two simple incentive mechanisms in a public goods experiment. J Pub Econ 92(1): 54-90.

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

Banks JS, Plott CR, Porter DP (1988) An experimental analysis of unanimity in public goods provision mechanisms. Rev Econ Stud 55(2): 301-322.

Blonski M, Spagnolo G (2015) Prisoners' other dilemma. Intl J Game Theory 44(1): 61-81.

Charness G, Fréchette GR, Qin C-Z (2007) Endogenous transfers in the prisoner's dilemma game: An experimental test of cooperation and coordination GEB 60(2): 287-306.

Cooper R, DeJong DV, Forsythe R, Ross TW (1996) Cooperation without reputation: experimental evidence from prisoner's dilemma games. GEB 12(2): 187-218.

Fehr E, Gächter S (2000) Cooperation and punishment in public goods experiments. Am Econ Rev 90(4): 980-994.

Fehr E, Powell M, Wilkening T (2014) Handing out guns at a knife fight: Behavioral limitations of subgame-perfect implementation, ECON – Working Paper 171, University of Zurich.

Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. Exper Econ 10(2): 171-178.

Forsythe R, Rietz T, Myerson R, Weber R (1996) An experimental study of voting rules and polls in three-candidate elections. Intl J Game Theory 25(3): 355-383.

Guala F (2012) Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. Behav Brain Sci 35(01): 1-15.

Hamaguchi Y, Mitani S, Saijo T (2003) Does the Varian mechanism work?-Emissions trading as an example. Intl J Bus Econ 2(2), 85-96.

Huang X, Masuda T, Okano Y, Saijo T (2016). Cooperation among behaviorally heterogeneous players in social dilemma with stay or leave decisions. KIER discussion paper series no. 944, Kyoto University.

Kalai E (1981) Preplay negotiations and the prisoner's dilemma. Math Soc Sci 1(4): 375-379.

Kimbrough EO, Sheremeta RM (2013) Side-payments and the costs of conflict. Intl J Ind Organ 31: 278-286.

Masuda T, Okano Y, Saijo T (2014) The minimum approval mechanism implements the efficient public good allocation theoretically and experimentally. Games Econ Behav 83: 73-85.

Midler E, Figuières C, Willinger M (2015) Choice overload, coordination and inequality: Three hurdles to the effectiveness of the compensation mechanism? Soc Choice Welf 1-23.

Plott CR, Smith VL (2008) Handbooks in economics 28: Handbook of experimental economics results. Elsevier, Amsterdam.

Qin, CZ (2005) Penalties and rewards as inducements to cooperate. Department of Economics, University of California-Santa Barbara.

Saijo T, Okano Y, Yamakawa T (2016) The approval mechanism experiment: A solution to prisoner's dilemma. KUT-SDES Working Paper no. 2015-12 (Revised), Kochi University of Technology.

Saijo T, Shen J (2015) Mate choice mechanism for solving a quasi-dilemma RIEB Discussion Paper no. 2015-34, Kobe University.

Selten R (1975) Reexamination of the perfectness concept for equilibrium points in extensive games. Intl J Game Theory 4(1): 25-55.

Selten R (1991) Properties of a measure of predictive success. Math Soc Sci 21(2): 153-167.

Varian HR (1994) A solution to the problem of externalities when agents are well-informed. Am Econ Rev 84(5): 1278-1293.