KOCHI UNIVERSITY OF TECHNOLOGY

Social Design Engineering Series *SDES-2017-22*

Mate Choice Mechanism for Solving a Quasi-Dilemma

Tatsuyoshi Saijo
School of Economics and Management, Kochi University of Technology
Research Center for Future Design, Kochi University of Technology

Junyi Shen
Research Institute for Economics and Business Administration, Kobe University

30th October, 2017

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

Mate Choice Mechanism for Solving a Quasi-Dilemma

Tatsuyoshi Saijo
Research Center for Future Design, Kochi University of Technology
Junyi Shen*
Research Institute for Economics and Business Administration, Kobe University

Abstract

Saijo, Okano, and Yamakawa (2015) showed that the mate choice mechanism for a symmetric prisoner's dilemma (*PD*) game implements cooperation in backward elimination of weakly dominated strategies (*BEWDS*), and it attained almost full cooperation in their experiment. This study theoretically shows, first, that this mechanism works well in the class of quasi-dilemma (*QD*) games, such as asymmetric *PD* games and coordination games. Second, the class of *BEWDS*-implementable games is exactly the same as the class of *QD* games. Third, the mechanism cannot implement cooperation in a subgame perfect equilibrium. Finally, we confirm that the mate choice mechanism works well experimentally for an asymmetric *PD* game.

Keywords: Asymmetric prisoner's dilemma; Quasi-dilemma; Mate choice mechanism; Cooperation; Experiment
JEL codes: C72, C92, D74, P43

* Corresponding author: Junyi Shen. 2-1 Rokkodai, Kobe, Hyogo 657-8501, Research Institute for Economics and Business Administration, Kobe University, Japan. Tel: +81 78 803 7013. Email: shen@rieb.kobe-u.ac.jp

1. Introduction

Over the past six decades, the prisoner's dilemma (*PD*) game has been extensively discussed in both the public and academic press (see Poundstone (2011) for a historical perspective). The *PD* game presents a scenario in which the outcome of one person's decision is determined by the simultaneous decisions of the other participants, resulting in a bad outcome for all of them (a Pareto-inefficient Nash equilibrium) if all act in their own self-interest. The key characteristic of this game is that while there are substantial gains that could be attained through cooperation, non-cooperation is dominant for each player (see Kuhn (2014) for an overview of PD literature).

Since participants in laboratory experiments consider non-monetary factors such as social norms and anonymity as well as monetary stakes, the cooperation rates, that is, the ratio of participants who chose cooperation, in PD game experiments are well above zero (see Chaudhuri (2011) for a survey), but not close to one. In order to increase the cooperation rates in PD games, researchers, such as Yamagishi (1986), Banks, Plott, and Porter (1988), Varian (1994), Fehr and Gächter (1999), Andreoni and Varian (1999), and Charness, Fréchette, and Qin (2007), started adding a stage *before* or *after* the PD games.[1] Adding a stage after the PD game allows participants to punish non-cooperation. Although participants in such games do not have a monetary incentive to punish other players, cooperation rates in experiments with punishment increase, but still do not become close to one (see Yamagichi (1986) and Fehr and Gächter (1999)). Varian (1994) introduced a compensation mechanism *before* the dilemma game. Under this mechanism each player is asked in the first stage to choose how much to pay his or her counterpart for cooperating. After learning the payments offered in the first stage, the players then play a normal *PD* game. Andreoni and Varian (1999) conducted experiments with the compensation mechanism and observed that cooperation rates increased from 25.8% when transfer payments were not feasible to 50.5% when transfer payments were permitted, and the cooperation rate was around 20% in the first two rounds. Charness, Fréchette, and Qin (2007) conducted experiments and observed that cooperation rates increased from 11–18% when transfer payments were not feasible to 43–68% when transfer payments were permitted.

Our aim in this study is to find one of the *simplest* possible mechanisms to solve social dilemmas, including the prisoner's dilemma, both experimentally and theoretically. Experimentally, we focus on designing a mechanism that can attain a Pareto efficient outcome

[1] The choice of a participant in the first stage is scrutinized by the other participant before deciding the choice in the second stage. As Levitt and List (2007) suggested, this might increase cooperation.

in a few rounds,[2] because we cannot repeat the same mechanism many times in real-life settings. Theoretically, we do not stick to Nash or Nash-type equilibrium concepts, but search for a behavioral principle among subjects in experiments that implements cooperation. In addition, we do not use punishment or reward to balance the budget, that is, there is no monetary inflow or outflow in the mechanism design.[3]

The study by Saijo, Okano, and Yamakawa (2015) is one of the first attempts to design such a mechanism for the *PD*.[4] They proposed the mate choice (*MC*) mechanism, which occurs *after* a *symmetric PD* game. After observing the choice of cooperation (*C*) or defection (*D*) in the *PD* game, each player is asked to approve or disapprove of the other's choice. If both approve it, the outcome is what they chose in the *PD* game; if at least one player disapproves of the other's choice, the outcome is the same as when both defect in the *PD* game.[5] Experimentally, Saijo et al. (2015) observed that the cooperation rate with the mechanism was 95.0% in round 1 and 96.9% over 19 rounds, when each subject was never matched with the same subject again in all rounds.[6] The (*C*, *C*) share, that is, the ratio of pairs in which both chose cooperation, was 90.0% in round 1 and 94.0% over 19 rounds. They also found that subjects' behavior was consistent with backward elimination of weakly dominated strategies (*BEWDS*) rather than Nash equilibrium (*NE*) or subgame perfect equilibrium (*SPE*) behavior. *BEWDS* is a procedure that eliminates weakly dominated strategies in each subgame, backwardly. The strategies that survive through the procedure are called *BEWDS* strategies. Theoretically, Saijo et al. (2015) proved that the *MC* mechanism implements cooperation in *BEWDS* for symmetric *PD* games.[7]

In two related studies, Masuda, Okano, and Saijo (2014) constructed a minimum approval mechanism, which is a version of the MC mechanism, in a public good economy when the number of players is two and their preferences are linear. They showed

---

[2] Chen (2005), for example, found that the stability property of mechanisms depends on their supermodularity. Supermodular mechanisms may require many periods to converge to a desired outcome. The goal of the endeavor is not to find such mechanisms but to find mechanisms that can attain a desired outcome in a few periods.

[3] According to Guala (2013), *strong reciprocity*, in which a player punishes other players using the player's own resources, is rare in human history.

[4] Saijo, Masuda, Okano and Yamakawa (2017) is a simplified version of Saijo et al. (2015).

[5] Because the *MC* mechanism does not have devices such as punishment or reward, it is budget balanced.

[6] This is called complete stranger matching, and only a few experiments employ this matching. Saijo et al. (2015) chose this matching since it is the least favorable design for cooperation with respect to matching.

[7] The *MC* mechanism uses unanimity. Banks, Plott, and Porter (1988) introduced a voting stage after a public good provision stage and observed that unanimity reduced efficiency. Researchers stopped pursuing this avenue after Banks et al. (1988) presented their findings. Furthermore, Masuda, Okano, and Saijo (2014) found that the *MC* mechanism cannot implement a Pareto-efficient allocation in *BEWDS* for an economy with a public good.

experimentally that the mean contributions ranged from 76.9% to almost 100.0%, with an average of 94.9%. Huang, Masuda, and Saijo (2014) constructed a simplified approval mechanism, which is also based on the MC mechanism, for a symmetric *PD* game in which the number of players was expanded from two to three. They observed experimentally that the mean cooperation rate increased from 44.4% in round 1 to above 90.0% in round 5 and maintained that level in the remaining 10 rounds.

From the above-mentioned studies, it seems that the MC mechanism performs very well in stimulating the players to cooperate in a *symmetric* environment. Consequently, the question of whether the MC mechanism is also effective in an *asymmetric* environment is natural. As Andreoni and Varian (1999) found, a participant with a relatively low payoff tends not to cooperate and this might influence the cooperation rate. Hence, in this paper we expand the domain of this mechanism from *symmetric PD* games to *asymmetric* games that are not necessarily *PD* games. We find theoretically that the *MC* mechanism implements cooperation in *BEWDS* for the class of *quasi-dilemma* (*QD*) games, which contains coordination games, including the stag hunt game and *PD* games. Furthermore, under several mild conditions, we show that the class of games implementing cooperation in *BEWDS* is exactly the same as the class of *QD* games, and that the *MC* mechanism cannot implement cooperation in *SPE*.

In order to test the performance of the MC mechanism experimentally in an asymmetric environment, we choose an asymmetric parameterization of the *PD* game ("Game 3" in Charness et al. (2007)) because the cooperation rate of this parameterization (42.9%) was worse than those for the other two asymmetric parameterizations (53.9% in "Game 1" and 68.1% in "Game 2"). This fact also motivates us to investigate whether the MC mechanism performs better than the compensation mechanism does. It should be noted that the compensation mechanism does not cover all *PD* games; in contrast, the MC mechanism covers all *PD* games *and* non-*PD* games. That is, there is a class of *PD* games in which the compensation mechanism cannot implement cooperation in *SPE*. Game 3 belongs to this class.

Experimentally, we observed that the cooperation rate with the MC mechanism in an asymmetric environment started at about 76.7% in round 1, rose to 86.7% in round 2, 93.3% in rounds 3 and 4, 96.7% in round 5, and then stayed above 98.0% in the remaining 14 rounds. The overall average cooperation rate over 19 rounds was 96.7%. The (*C*,*C*) share started at 56.7% in round 1, rose to 73.3% in round 2, to 86.7% in rounds 3 and 4, to 93.3% in round 5, and then stayed above 96.0% in the remaining 14 rounds. The overall average (*C*,*C*) share over 19 rounds was 93.5%. That is, the MC mechanism works reasonably well, although it took a few

rounds to achieve high (*C*,*C*) share in an asymmetric *PD* game.

A good example of the mate choice mechanism in an *asymmetric* environment is so-called *MAD* (Mutually Assured Destruction) that led the earth to the avoidance of nuclear disaster around the last half of the twentieth century. Even though superpower *A* attacks the other superpower *S* using nuclear weapons, superpower *S* can monitor the attack and then has enough time to mount the counterattack. In other words, this is a two-stage game where the first stage is a *PD* game, and the second stage is a special case of the approval stage. The approval in the second stage is "No (Further) Attack" and the non-approval is "(Counter) Attack." If a superpower decides to choose "Attack" in the *PD* game, she must choose "No (Further) Attack" automatically in the second stage since she has already chosen "Attack" in the first stage. Then each chooses "No Attack" or "No Action" in the first stage, and then chooses "No (Further) Attack" in the second stage is the unique *BEWDS* path.[8] Notice that the second stage mechanism is not by man made one such as convention, but by an evolving mechanism due to technological constraints including the monitoring accuracy and the time lag between the discharge and explosion that are called second-strike capability by Russett, Starr and Kinsella (page 237, 2009). The technological progresses were due to the battle of holding hegemony over the other superpower.

There are many other examples of the mate choice mechanism. Consider a merger or a joint project of two companies. They must propose plans (the contents of cooperation) in the first stage, and then each faces the approval decision in the second stage. In order to resolve the conflicts such as prisoner's dilemma, interested parties usually form a committee consisting of representatives of the parties. Consider two companies facing confrontation on the standardizations of some product. Each company chooses cooperation (or compromise) or defection (or advocating of the own standard), and then the committee consisting of two company members and/or bureaucrats gives the approval. Another example is the two party system. Each party chooses either cooperation (or compromise) or defection (or insistence of policy for the own party), and then diet (or national assembly) plays a role of approval. The bicameral system also has two stages. One chamber decides a policy (or compromise) and the other chamber plays a role of approval. The negotiation process at United Nations also has this structure. Negotiators among relevant countries get together to find compromise, i.e., the

[8] Robert J. Aumann (2006) in his Nobel Lecture described *MAD* as an outcome of infinitely repeated games in order to maintain cooperation. The idea of approval mechanism is not to use infinite periods but to consider the game in two stages. Notice also that (*Attack*, *Attack*) is a part of *SPE*, but not a part of the *BEWDS* path. That is, it was fortunate that the decision makers of the superpowers did not follow this path.

content of cooperation in the first week and then high ranked officials such as presidents and prime ministers get together to approve or disapprove it in the second week. Adding the second stage in resolving conflicts has been used widely in our societies.

The paper is organized as follows. Section 2 describes the MC mechanism applied to *QD* games. Section 3 proves that *BEWDS* implementable games are *QD* games and shows that the *MC* mechanism cannot implement cooperation in *SPE*. Section 4 presents the experimental design, and Section 5 presents the results. Section 6 provides suggestions for further research.

2. The mate choice mechanism and quasi-dilemma games

Consider a 2 × 2 game that has two strategies: cooperation (*C*) and defection (*D*).

		Player 2	
		C	*D*
Player 1	*C*	$(a,v)=V$	$(b,w)=W$
	D	$(c,x)=X$	$(d,z)=Z$

Figure 1. A *QD* game.

Define a payoff function p for the game as follows: $p(C,C) = (a,v) = V$, $p(D,C) = (c,x) = X$, $p(C,D) = (b,w) = W$, and $p(D,D) = (d,z) = Z$. If p satisfies $V > Z$ ($a > d$ and $v > z$), $X \not\geq Z$ ($d > c$ or $z > x$), and $W \not\geq Z$ ($d > b$ or $z > w$), then p is a *QD game*; if p satisfies $V > Z$, $(c,d) > (a,b)$ and $(w,z) > (v,x)$, then p is a *PD game*.[9] Coordination games, including the stag hunt game, are *QD* games.

Property 1. *A prisoner's dilemma game is a quasi-dilemma game.*

Proof. Let p be the payoff function of a *PD* game. Then $(a,v) > (d,z)$, $(c,d) > (a,b)$, and $(w,z) > (v,x)$. Given $V = (a,v)$ and $Z = (d,z)$, let $X = (c,x)$ and $W = (b,w)$ satisfy the conditions. Thus, $c > a$, $z > x$, $d > b$, and $w > v$. Since $z > x$ and $d > b$, it follows that ($d > c$ or $z > x$) and ($d > b$ or $z > w$). That is, $X \not\geq Z$ and $W \not\geq Z$. ∎

[9] ">" indicates that each element of the left-hand-side vector is "strictly greater than" each element of the right-hand-side vector, and "≥" indicates that each element of the left-hand-side vector is "greater than or equal to" each element of the right-hand-side vector.

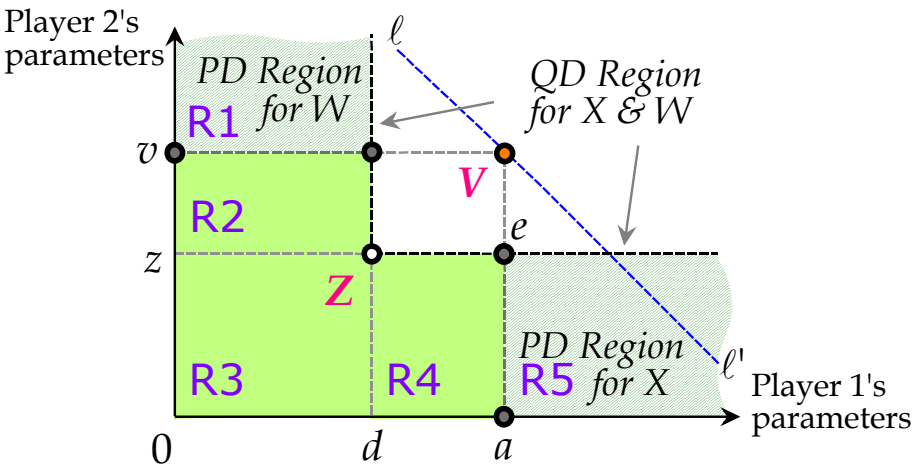

Figure 2. Possible locations of *X* and *W* in *QD* and *PD* games.

Figure 2 shows the possible locations of *X* and *W* in *QD* and *PD* games. The region for *QD* games is the dark *L*-shaped area (i.e., R1, R2, R3, R4, and R5), and the regions for *PD* games are R1 for *W* and R5 for *X*. The line $\ell - \ell'$ shows points where the sum of benefits to both players is equal to $a + v$. If *V* must be Pareto efficient, then *X* and *W* must lie under $\ell - \ell'$. We do not impose this condition hereafter.

Let us define the *MC* mechanism. Players first choose either *C* or *D* in Figure 1. After this stage, each subject can approve (*y*) or disapprove (*n*) of the other player's choice in the second stage. If both approve the other player's choice, the outcome is what they chose; if either one disapproves of the other player's choice, the outcome is that both defect. Let M_i, m_i, and u_i be player *i*'s choice between *C* and *D*, *i*'s choice between *y* and *n*, and *i*'s payoff, respectively. Then, the *MC* mechanism is defined by the following rule: if $m_1 = m_2 = y$, then $(u_1, u_2) = p(M_1, M_2)$; otherwise, $(u_1, u_2) = p(D, D)$. In general, the two-stage game without any rule specification is called an *approval* mechanism.

Saijo et al. (2015) considered five possible behavioral principles for the *MC* mechanism in their experiments: the *NE*, the *SPE*, evolutionarily stable strategies (*ESS*), neutrally stable strategies (*NSS*), and *BEWDS*. Of the five behavioral principles, the data from their experiments are most compatible with *BEWDS*.

Since we focus upon *SPE* and *BEWDS*, let us define them. As we show later, since there is no need to consider mixed strategies in our framework, a profile of strategies indicates

an assignment of a pure strategy for each information set. A profile of strategies is an *SPE* if the restriction of the profile at each subgame is a Nash equilibrium. Let us fix a subgame, and we say that a strategy at an information set in the game *survives the elimination of weakly dominated strategies* if the strategy is not weakly dominated by any other strategies available at that information set.[10] A profile of strategies is a *BEWDS* if the restriction of the profile at each subgame survives the elimination of weakly dominated strategies. A mechanism *implements* cooperation in *SPE* (or *BEWDS*) if all players choose cooperation in the first stage under *SPE* (or *BEWDS*). Next, we show that the *MC* mechanism with a *QD* game implements cooperation under *BEWDS*.

Property 2. *The mate choice mechanism with a QD game implements the cooperative outcome under BEWDS.*

Proof and Interpretation. The *MC* mechanism with a *QD game* has the four subgames shown in Figure 3. Because of the definition of the *MC* mechanism, we have $(u_1, u_2) = (d, z)$ for (*y,n*), (*n,y*), and (*n,n*) in each subgame. This was termed the *MC flat* by Saijo et al. (2015), because the three cells other than (*y,y*) have the same payoff vector.

Consider subgame *CC*, where both players have chosen *C*. Player 1 must compare (*a,d*) and (*d,d*). Because $a > d$, player 1 chooses *y* because of the elimination of weakly dominated strategies. However, player 1 does not necessarily compare two vectors because of the *MC* flat. Player 1 should compare *a* and *d* to understand this domination.

Although player 1 can choose either *y* or *n*, basing this on the elimination of weakly dominated strategies alone, player 1 must additionally consider player 2's choice. Player 1 compares *v* and *z*, and hence, player 1 understands that player 2 chooses *y*. Thus, player 1 understands that the choice at (*C,C*) is (*y,y*), which is shown by the bold square in Figure 3 for subgame *CC*. Therefore, player 1 can fill the (*C,C*) part of the reduced normal form game with (*a,v*), which is located above the four subgames in Figure 3.

Consider subgame *DC*. Because $X \not\geq Z$, it must be that $d > c$ or $z > x$. Suppose that $z > x$. Then, player 2 chooses *n* and understands that the outcome is (*d,z*) regardless of the choice of player 1 in this subgame. However, as a thought experiment, player 1 must consider player 2's choice if it were the case that $c > d$. That is, although player 1 would choose *y*, player 1 could

[10] For the game shown in Figure 1, *C weakly dominates D* for player 1 if and only if $a \geq c$, $b \geq d$, and there is at least one strict inequality. Notice that no strategies could survive if we use strong domination with strict inequality of each element.

not identify which of (c,x) and (d,z) would be realized without knowing player 2's choice. Suppose $d > c$. Player 1 chooses n and understands that the outcome is (d,z) without considering player 2's choice.

Repeating the same procedure at each of the subgames CD and DD, player 1 can construct the reduced normal form game above the four subgames in Figure 3. Because the game also has the MC flat and $a > d$, player 1 chooses C.

If player 1 understands that player 2's position is the same as that of player 1 using the same procedure, player 1 is convinced that player 2 also chooses (C,y). If so, the equilibrium path under $BEWDS$ is $((C,C),(y,y))$. Thus, we simply write (C,C,y,y) hereafter. ∎

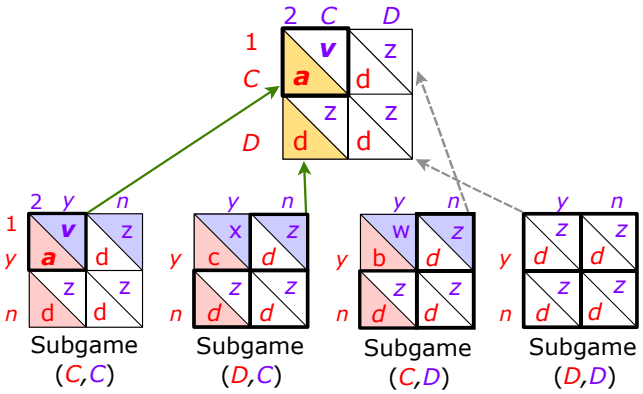

Figure 3. Construction of the reduced normal form game.

Although this proof is mathematically almost the same as that of Saijo et al. (2015), there are several differences between the two because of possible asymmetry. First, because they used the symmetric payoff table, $d = z$, all components of the three cells have the same number. In this sense, these three cells have the *completely* flat property. Therefore, the change from $d = z$ to $d \neq z$ might increase the player's burden to understand the game. Second, (a,d) is different from (v,z), whereas $(a,d) = (v,z)$ in Saijo et al. (2015). This change also increases the players' burden. The same argument can be applied to all subgames. Third, players who understand the strategic implications of vectors V, W, X, and Z in the symmetric case might not be able to understand them in the asymmetric case under $BEWDS$. Finally, even though players may understand these points, the difference between a and v may evoke an "equity" module in their brains, which could trigger "non-rational" motivation.

An important property of the MC mechanism is that it is *onto*: the set of possible

payoffs of a *QD* game is equal to the set of those of the *MC* mechanism. The former is $\{V,W,X,Z\}$, and the latter must be the same. For example, although player 1 does not want to choose y at (C,D), W is the payoff pair if both choose y. This condition excludes any payoff flow from or to the game. In other words, no outside penalty or reward is given in order to maintain a balanced budget.[11]

3. Games that are implementable by *BEWDS* are quasi-dilemma games

We consider the tightness of the parameter space of *QD game*s and the implementable parameter space of *BEWDS*. If the latter is larger, the idea of *BEWDS* implementation can be applied to many other 2 × 2 games. However, we show that these two spaces are identical under several assumptions.[12]

First, the approval mechanism satisfies *forthrightness*: if both choose y in the second stage after the choice of a strategy pair in the first stage, the outcome must be that strategy pair.[13] That is, the outcome must be what they choose whenever both choose y. Second, the approval mechanism has a *flat*: the outcome of the second stage, except for (y,y), and that of the reduced game, except for (C,C), are the same. Thus, we have the following property.

Property 3. *Suppose that $V > Z$ and that an approval mechanism satisfies forthrightness with the flat, Z. Then, the class of games implementing cooperation in BEWDS is exactly the same as that of QD games.*

Proof. Consider any approval mechanism implementing outcome V in *BEWDS*. Because forthrightness is satisfied, if both choose y, the payoffs of subgames CC, CD, DC, and DD must be V, W, X, and Z, respectively. Because Z is the flat and the mechanism implements cooperation, the payoff should be Z in subgames CD and DC. That is, $(d > b$ or $z > w)$ and $(d > c$ or $z > x)$ for subgames CD and DC, respectively. Because $\overline{(d > b \text{ or } z > w)} = (d \le b \text{ and } z \le w) = (W \ge Z)$, $(d > b$ or $z > w)$ is equivalent to $W \not\ge Z$. Similarly, we obtain $X \not\ge Z$. That is, the class of games implementing cooperation in *BEWDS* is exactly the same as the class

[11] The money-back-guarantee mechanism introduced by Dawes, Orbell, Simmons, and van de Kragt (1986) and Isaac, Schmidtz, and Walker (1989) is not *onto*. Let C be a fixed amount of contribution for public good provision, and let D be no contribution. The money-back-guarantee mechanism returns the contribution to a player if both do not contribute, and hence, X and W cannot be payoffs produced by the mechanism. If player 1 is a utilitarian, and hence prefers X to Z, the *MC* mechanism can lead to X unlike the money-back-guarantee mechanism.

[12] These assumptions are essentially the same as those introduced by Saijo et al. (2015).

[13] Saijo, Tatamitani, and Yamato (1996) introduced forthrightness in natural mechanism design.

of *QD* games. ■

Let us now consider the *SPE*s of *QD* games. Since the Nash equilibrium payoffs are (*a,v*) and (*d,z*) at subgame *CC* and (*d,z*) at subgames *DC, CD*, and *DD*, two cases exist in the reduced normal form games shown in Figure 4. Because all combinations, (*C,C*), (*D,C*), (*C,D*), and (*D,D*), are Nash equilibria of the games, they are also the outcomes of *SPE*s. Therefore, the *MC* mechanism cannot implement cooperation in *SPE*. That is, multiple Nash equilibria in a subgame generate multiple *SPE*s due to the *MC* flat. Furthermore, players in an *SPE* must consider how the other player behaves in each subgame.

		Player 2	
		C	*D*
Player 1	*C*	***a,v***	*d,z*
	D	*d,z*	***d,z***

(i) (*y,y*) in subgame *CC*

		Player 2	
		C	*D*
Player 1	*C*	***d,z***	***d,z***
	D	***d,z***	***d,z***

(ii) (*n,n*) in subgame *CC*

Bold italic cells indicate Nash Equilibria in the reduced normal form games.

Figure 4. Two reduced normal form games.

Consider now informational burden to a player using *BEWDS*. Since subjects can instantaneously understand that each subgame in the second stage has the *MC* flat, and can identify (*d,z*) as the outcome of subgame *DD* in Figure 3, the cells or triangles that subjects must see or consider are dark parts in subgames *CC, CD* and *DC* in the second stage under *BEWDS*. Subject 1's own possible outcomes to be compared are the two lower left triangles of the left column. Hence, the number of triangles that each subject must see is four in each of subgames *CC, CD* and *DC*. That is, the remaining triangles are unnecessary for their decision making including the lower right cell.

Let us consider the *minimum* informational or intellectual requirement to achieve *CCyy* under *BEWDS*. Consider subject 1. The information of the two lower left triangles of the left column in each of subgames *CC, CD*, and *DD* is enough to solve or choose either *y* or *n* in each subgame, but this is not enough to solve the two stage game since subject 1 cannot identify which cell would be realized without having the information of the two upper right triangles of the upper row at subgames *CC* and *DC*. In other words, subject 1 must use *theory of mind* to understand which strategy subject 2 chooses. If this is successful, subject 1 can construct the reduced normal form game out of two stages shown at the top in Figure 3.

During this construction, subject 1 understands that (s)he really needs to know for the decision making is the two lower left triangles of the left column in the reduced normal form game, i.e., subject 1's outcomes of subgames *CC* and *DC*. Using this information, subject 1 chooses *C*. In this sense, subject 1 must have *backwardability* that identifies chosen cells in subgames *CC, CD* and *DD*, and then finds his or her own two triangles corresponding to subgames *CC* and *DC* in the reduced normal form game. Finally, subject 1 also uses a simple heuristic: "*the other subject thinks the same way as I think*." For example, subject 1 who understands the outcome of subgame *DC* can find the outcome of subgame *CD* using this heuristic. These two simplified methods mitigate subjects' burden considerably.[14,15] Of course, these facts do not necessarily support *BEWDS* in general since the facts are specific to the *MC* mechanism.

In the next two sections, we describe our experimental investigation of the *MC* mechanism and its results.

4. Experimental design

Our experimental focus is payoff asymmetry in the *PD* game. We conducted experiments using an asymmetric *PD* game with the *MC* mechanism (*AsymPDMC*). For comparison with the data for *AsymPDMC*, we borrowed the data for a symmetric *PD* game with the *MC* mechanism (*SymPDMC*) and a symmetric *PD* game without the *MC* mechanism (*SymPD*) from Saijo et al. (2015).

We chose an asymmetric payoff table in which cooperation cannot be implemented by the compensation mechanism in *SPE*, but can be implemented by the *MC* mechanism in *BEWDS*. The compensation mechanism also has two stages. It asks players to transfer money to the other player in the first stage, and then, both play a *PD* game. The monetary transfer

---

[14] Player 1 must compare two numbers six times in order to decide (*C*,*y*) in Figure 3: two comparisons (i.e., my own *a* and *d*, and the other's *v* and *z*) in subgame *CC*, one comparison (i.e., my own *d* and *c*, and the other's choice does not matter since my own outcome is *d* regardless the choice of the other) in subgame *CD*, two comparisons in subgame *DC*, and one comparison between *a* and *d* in the reduced normal form game. This is quite a contrast when we find Nash equilibria of the two stage game. Since the number of information set is 5, each player has 2^5 strategies. Player 1 must compare (2^5-1) numbers to find best responses for any given strategies of player 2. This indicates that player 1 must compare two numbers $(2^5-1)\times 2^5$ times. In order to find the Nash equilibria, player 1 must find the best responses of player 2, and hence must compare two numbers $(2^5-1)\times 2^5$ times. That is, the number of comparisons is $2\times(2^5-1)\times 2^5 = 1984$ which is more than 300 times of 6, which might trigger qualitative difference between *NE* and *BEWDS*. See Saijo et al. (2015) for detailed comparison of several equilibrium concepts such as Nash equilibria, subgame perfect equilibria, neutrally stable strategies, and evolutionarily stable strategies for symmetric games.

[15] A reviewer suggested to introduce trembling-hand SPE or conditional cooperator in order to eliminate "bad" outcomes, but these are our future agenda.

must be done when the other player chooses cooperation in the *PD* stage. Varian (1994) designed the compensation mechanism in a general setting, and then Andreoni and Varian (1999) and Charness et al. (2007) conducted experiments using *PD* games. As Charness et al. (2007) showed, the compensation mechanism does not cover the entire class of *PD* games. In this sense, we chose the least favorable matrix (i.e., the Game 3 matrix in Charness et al. (2007)) to examine cooperation levels in our experimental design. Figure 5 shows the symmetric and asymmetric payoff matrices.[16] The symmetric payoff table comes from Saijo et al. (2015) and the asymmetric payoff matrix comes from Charness et al. (2007).

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

(i) Symmetric payoff matrix

		Player 2 *C*	Player 2 *D*
Player 1	*C*	44,36	8,44
	D	52,0	32,28

(ii) Asymmetric payoff matrix

Figure 5. Symmetric and asymmetric payoff matrices.

All the above-mentioned experiments were carried out in Osaka University during the period from November 2009 to December 2011. The *AsymPDMC* and *SymPDMC* experiments each had three sessions, and the *SymPD* experiment had one session. Twenty subjects participated in each session and no subject attended more than one session. We recruited these 140 undergraduate subjects with different majors through campus-wide advertisements. They were told that there would be an opportunity to earn money in a research experiment.

At the beginning of the experiment, each subject was given a set of printed instructions and a record sheet. Instructions were read aloud by an experimenter. After that, subjects were given five minutes to ask private questions. Communication among subjects was prohibited, and we declared that the experiment would be stopped if it was observed. This never happened. There was no practice period. We used the z-Tree software (Fischbacher, 2007) for the experiment.

The experimental procedure was as follows. We assigned the 20 subjects seated at computer terminals in each session to 10 pairs. These pairings were anonymous and

[16] In the experiment, we used payoff numbers that are 100 times the numbers in Figure 5 due to the exchange rate between the Japanese yen and the US dollar.

determined in advance in order not to pair the same two subjects more than once. Since most previous studies, such as those of Andreoni and Varian (1999) and Charness et al. (2007), have employed random matching among four to eight subjects (two to four groups),[17] such repetition necessarily entails pairings of the same two subjects. Therefore, compared with previous experiments, this "complete" strangers design might reduce the possibility of cooperation among subjects.

Let us explain the *PDMC* experiment. When the period started, each subject selected either *A* (defection) or *B* (cooperation) in the choice (or *PD*) stage and then inputted the choice into a computer and also noted it on the record sheet. After that, each subject explained the reason behind this choice in a small box on the record sheet. [18] The next step was the decision (or approval) stage. Based on the knowledge of the other subject's choice, each subject chose to either "accept" or "reject" it and then inputted the decision into a computer, noted it on the record sheet, and explained his or her reasoning as before. Once subjects had finished the task, each 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. However, neither the choices nor the decisions in pairs other than "your" own were shown on the computer screen. This ended one period. The *PD* experiment omitted the approval stage. After finishing all 19 periods, every subject filled in a questionnaire.

Each session lasted approximately 90 minutes including the time spent on answering the post-experiment questionnaires and payment. Subjects earned, on average, 5233 JPY (about 43.61 USD, using 1 USD=120JPY), 4873 JPY (about 40.61 USD), and 3920 JPY (about 32.67USD) in *AsymPDMC*, *SymPDMC*, and *SymPD* sessions, respectively.

5. Experimental results

Figure 6 shows the time paths of cooperation rates over the 19 periods. The cooperation rate in each period is defined as the ratio of number of subjects choosing *C* to the total number of subjects. As shown, the cooperation rate of *PD* started at 15% in the first three periods, and then ranged between 5% and 10% in the next 16 periods. In contrast, the cooperation rate of *SymPDMC* was always above 90%, while that of *AsymPDMC* started at about 76.7% in the first period, rose to 86.7% in the second period, and then stayed above 90%

[17] Charness et al. (2007) divided 16 subjects in one session into four separate groups, with four subjects in each group interacting only with each other over the course of the session.

[18] We had wished that we could analyze these reasons. However, most of the subjects did not write down reasons or just wrote "in order to earn more money", which leaves analyses of this information uninformative. Hence, we do not report these results in the paper.

in the remaining 17 periods.

The large gap in cooperation rates between *PD* and *PDMC* (either symmetric or asymmetric) was statistically supported by the proportion test. All the p values for comparing the cooperation rate of *AsymPDMC* or *SymPDMC* with that of *PD* are smaller than 0.001 in each period, which suggests that introducing the second stage after the *PD* game dramatically increases the cooperation rate in both symmetric and asymmetric *PD* games.

For the comparison of *AsymPDMC* with *SymPDMC*, we ran the proportion test using both the data from each period and the data pooled over all periods. The two-tailed p values are reported in Table 1 (see the fourth column from the left). Generally, there is no significant difference in cooperation rates between asymmetric *PD* and symmetric *PD* games when the data were pooled over all periods ($p = 0.7212$). In the first two periods, the cooperation rate is significantly lower in the asymmetric environment. Thereafter, the cooperation rate in the asymmetric MC mechanism is not significantly lower than its symmetric counterpart. In fact, a significantly higher cooperation rate in *AsymPDMC* is observed in 3 periods out of the remaining 17 periods (specifically, periods 7, 14, and 15). In addition, Andreoni and Varian (1999) found a significant difference in cooperation rates between subjects with a relatively low payoff (41.8%) and their counterparts with a relatively high payoff (59.2%). However, as shown in Table 2 we did not find such significant differences in cooperation rates between these two types of subject. Moreover, compared to the cooperation rate with the compensation mechanism in Charness et al. (2007) the one in our *AsymPDMC* is obviously higher, which suggests the superiority of an MC mechanism over a compensation mechanism.

With regard to the share of the (*C*,*C*) combination, Figure 7 shows its time paths over the 19 periods. Applying the proportion test, we found that the (*C*,*C*) share is significantly higher in either *AsymPDMC* or *SymPDMC* than in *PD* in each period (all p values < 0.001). Additionally, as indicated in the last column of Table 1, there is no significant difference in the share of the (*C*,*C*) combination between *AsymPDMC* and *SymPDMC* ($p = 0.6031$) when the data were pooled over all periods. However, if we look at the p values by period, we find that the (*C*,*C*) share is significantly lower in the asymmetric environment than in the symmetric environment in the first four periods. Thereafter, as with the cooperation rate, the (*C*,*C*) share of *AsymPDMC* is once again always statistically higher than or equal to that of its symmetric counterpart.

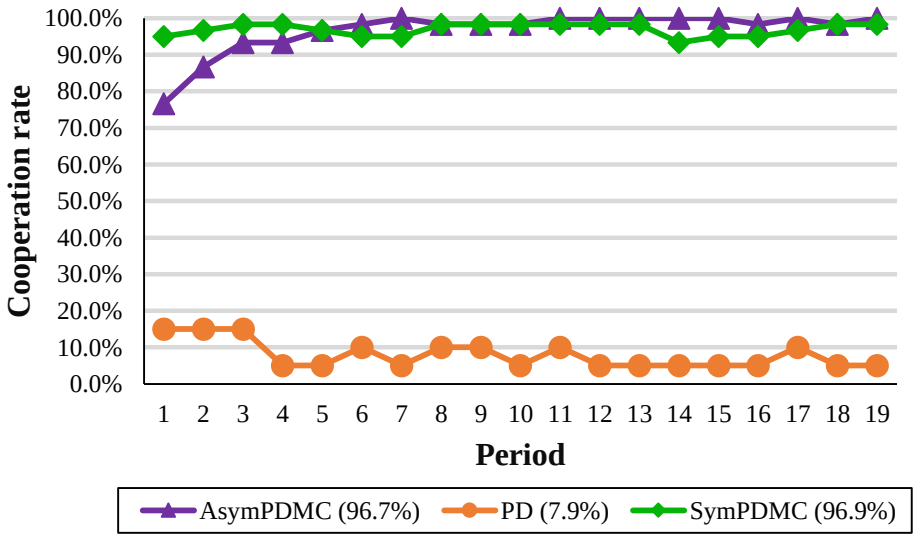

Figure 6. Cooperation rates by period.

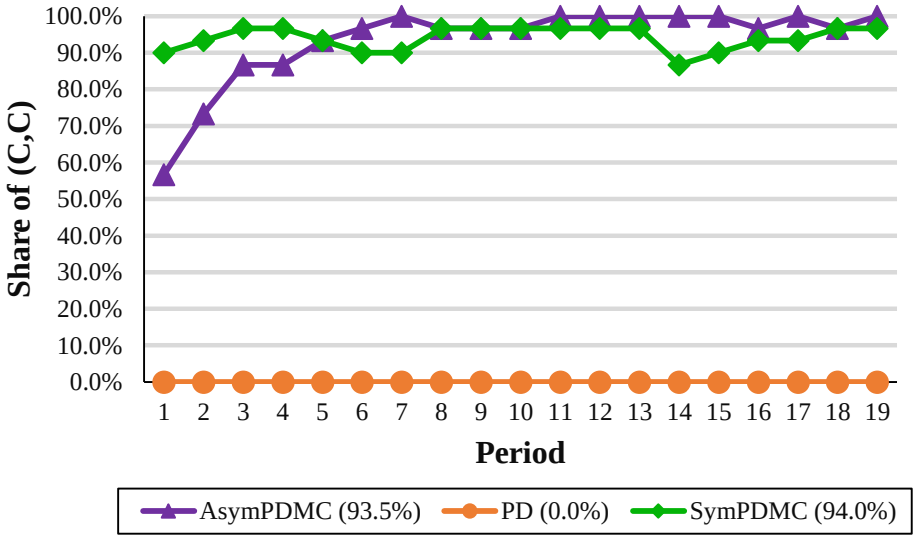

Figure 7. Share of (*C*,*C*) by period.

Table 1. Proportion test results for *AsymPDMC* vs. *SymPDMC*.

Period	Cooperation rate			Share of (*C*,*C*)		
	AsymPDMC	*SymPDMC*	*p* value	*AsymPDMC*	*SymPDMC*	*p* value
1	76.7%	95.0%	0.0040	56.7%	90.0%	0.0000
2	86.7%	96.7%	0.0475	73.3%	93.3%	0.0033
3	93.3%	98.3%	0.1705	86.7%	96.7%	0.0475
4	93.3%	98.3%	0.1705	86.7%	96.7%	0.0475
5	96.7%	96.7%	1.0000	93.3%	93.3%	1.0000
6	98.3%	95.0%	0.3091	96.7%	90.0%	0.1432
7	100.0%	95.0%	0.0794	100.0%	90.0%	0.0120
8	98.3%	98.3%	1.0000	96.7%	96.7%	1.0000
9	98.3%	98.3%	1.0000	96.7%	96.7%	1.0000
10	98.3%	98.3%	1.0000	96.7%	96.7%	1.0000
11	100.0%	98.3%	0.3153	100.0%	96.7%	0.1538
12	100.0%	98.3%	0.3153	100.0%	96.7%	0.1538
13	100.0%	98.3%	0.3153	100.0%	96.7%	0.1538
14	100.0%	93.3%	0.0419	100.0%	86.7%	0.0034
15	100.0%	95.0%	0.0794	100.0%	90.0%	0.0120
16	98.3%	95.0%	0.3091	96.7%	93.3%	0.4022
17	100.0%	96.7%	0.1538	100.0%	93.3%	0.0419
18	98.3%	98.3%	1.0000	96.7%	96.7%	1.0000
19	100.0%	98.3%	0.3153	100.0%	96.7%	0.1538
All periods	96.7%	96.9%	0.7212	93.5%	94.0%	0.6031

Notes: The reported *p* values are based on the two-tailed proportion test.

Table 2. Cooperation rates of subjects with low payoff and with high payoff in *AsymPDMC*.

Period	*Subjects with relative low payoff*	*Subjects with relative high payoff*	*p* value
1	80.0%	73.3%	0.5416
2	83.3%	90.0%	0.4475
3	93.3%	93.3%	1.0000
4	93.3%	93.3%	1.0000
5	100.0%	93.3%	0.1503
6	100.0%	96.7%	0.3132
7	100.0%	100.0%	–
8	96.7%	100.0%	0.3132
9	96.7%	100.0%	0.3132
10	96.7%	100.0%	0.3132
11	100.0%	100.0%	–
12	100.0%	100.0%	–
13	100.0%	100.0%	–
14	100.0%	100.0%	–
15	100.0%	100.0%	–
16	96.7%	100.0%	0.3132
17	100.0%	100.0%	–
18	100.0%	96.7%	0.3132
19	100.0%	100.0%	–
All periods	96.7%	96.7%	1.0000

Notes: The reported *p* values are based on the two-tailed proportion test. A "–" means the proportion test cannot be performed because the cooperation rates in both conditions are 100.0%

6. Concluding remarks

We have shown theoretically that the MC mechanism implements cooperation in *BEWDS* for *QD* games and that *BEWDS*-implementable games are *QD* games. *QD* games include not only *PD* games but also coordination games. Furthermore, the mechanism cannot implement cooperation in *SPE*.

The MC mechanism is essentially a unanimous voting rule for two players, and it can

be interpreted as a "minimum" communication device to achieve cooperation. In the first stage, each player reveals their choice of *C* or *D*. Then knowing the other's choice, each chooses *y* or *n*. If both choose *y*, the outcome is what they chose in the first stage; otherwise, the outcome is (*D*,*D*). This procedure can be a *natural* way to avoid conflict, such as a *PD* or a coordination situation, in daily life. Using functional near-infrared spectroscopy (*fNIRS*), Nagatsuka, Shinagawa, Okano, Kitamura, and Saijo (2013) found that compared with the *PD* game, subjects made their choices with less stress when using the MC mechanism. We also found experimentally that except for first few periods, the MC mechanism works well in an asymmetric environment.

Masuda et al. (2014) expanded the idea of the MC mechanism to public good provision and showed that the minimum approval mechanism implements an efficient allocation in *BEWDS* both theoretically and experimentally. Furthermore, Huang et al. (2014) designed a simplified approval mechanism in the spirit of the MC mechanism in a social dilemma and showed theoretically and experimentally that it implements cooperation in *BEWDS* when there are at least two players. However, designing a reasonable approval mechanism to implement cooperation with more than two choices and players is still an open question that needs to be answered.

References

Andreoni, J., and Varian, H. 1999. "Preplay Contracting in the Prisoners' Dilemma." *Proceedings of the National Academy of Sciences*, 96(19), 10933–10938.

Aumann, Robert J. 2006. "War and Peace." *Proceedings of the National Academy of Sciences,* 103(46): 17075-78.

Banks, J.S., C.R. Plott, and D.P. Porter. 1988. "An Experimental Analysis of Unanimity in Public Goods Provision Mechanisms." *Review of Economic Studies*, 55(2): 301–322.

Charness, G., G.R. Fréchette, and C.-Z. Qin. 2007. "Endogenous Transfers in the Prisoner's Dilemma Game: An Experimental Test of Cooperation and Coordination." *Games and Economic Behavior*, 60(2): 287–306.

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

Chen, Y. 2005. "Dynamic Stability of Nash-efficient Public Goods Mechanisms: Reconciling Theory and Experiments." In A. Rapoport and R. Zwick (eds.). *Experimental Business Research.* Vol II, 185–200.

Cooper, R., D.V. DeJong, R. Forsythe, and T.W. Ross. 1996. "Cooperation Without Reputation:

Experimental Evidence from Prisoner's Dilemma Games." *Games and Economic Behavior*, 12(2), 187–218.

Dawes, R., J. Orbell, R. Simmons, and A. van de Kragt. 1986. "Organizing Groups for Collective Action." *American Political Science Review,* 80(4), 1171–1185.

Fehr, E., & Gächter, S. (1999). Cooperation and punishment in public goods experiments. *Institute for Empirical Research in Economics working paper*, (10).

Fischbacher, U. 2007. "Z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10(2), 171–178.

Guala, F. 2013. "Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate." *Behavioral and Brain Sciences,* 35, 1–59.

Huang, X., T. Masuda, Y. Okano, and T. Saijo. 2014. "Cooperation among Behaviorally Heterogeneous Players in Social Dilemma with Stay or Leave Decisions." Working Paper in Social Design Engineering Series. SDES-2014-7, Research Center for Future Design, Kochi University of Technology.

Isaac, R.M., D. Schmidtz, and J.M. Walker. 1989. "The Assurance Problem in a Laboratory Market." *Public Choice,* 62 (3), 217–236.

Kuhn, S. 2014. "Prisoner's Dilemma", *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2014/entries/prisoner-dilemma/>.

Levitt, S. D., & List, J. A. (2007). "What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives*, 21(2), 153-174.

Masuda, T., Y. Okano, and T. Saijo. 2014. "The Minimum Approval Mechanism Implements the Efficient Public Good Allocation Theoretically and Experimentally." *Games and Economic Behavior,* 83, 73–85.

Nagatsuka, M., H. Shinagawa, Y. Okano, Y. Kitamura, and T. Saijo. 2013. "Using Economic Games to Investigate the Neural Substrates of Cognitive Processes." *American Journal of Clinical Medicine Research*, 1(4), 71–74.

Poundstone, W. (2011). *Prisoner's dilemma*. Anchor.

Russett, B., H. Starr and D. Kinsella. 2009. *World Politics: The Menu for Choice*, Wadsworth Publishing. 9th edition.

Saijo, T., T. Masuda, Y. Okano, and T. Yamakawa. 2017. "Approval Mechanism to Solve Prisoner's Dilemma," submitted to *Social Choice and Welfare*, revision requested.

Saijo, T., Y. Okano, and T. Yamakawa. 2015. "The Approval Mechanism Experiment: A Solution to the Prisoner's Dilemma." Working Paper in Social Design Engineering Series. SDES-2015-12, Research Center for Future Design, Kochi University of Technology.

Saijo, T., Y. Tatamitani, and T. Yamato. 1996. "Toward Natural Implementation." *International Economic Review,* 37(4), 949–980.

Varian, H.R. 1994. "A Solution to the Problem of Externalities When Agents are Well-informed." *American Economic Review*, 84, 1278–1293.

Yamagishi, T. (1986). "The Provision of a Sanctioning System as a Public Good." *Journal of Personality and social Psychology*, 51(1), 110-116.