

October 14, 2011
Not for circulation!

**The Approval Mechanism Experiment:
A Solution to Prisoner's Dilemma†**

January 2010/Revised October 2011

Tatsuyoshi Saijo*[#], Yoshitaka Okano* and Takafumi Yamakawa*
Osaka University* and UCLA[#]

Abstract

Players can approve or reject the other choice of the strategy after announcing the choices in a prisoner's dilemma game. If both approve the other choice, the outcome is what they choose, and if either one rejects the other choice, it is the outcome when both defect, which is called the mate choice mechanism. The Nash equilibria (*NE*) and subgame perfect equilibria (*SPE*) of this two stage game have all possible combinations of cooperation and defection. However, the outcome of neutrally stable strategies (*NSS*) and backward elimination of weakly dominated strategies (*BEWDS*) is that both are cooperative. Furthermore, reciprocal behavior (if you cooperate, I will approve it) reinforces cooperation under *BEWDS*. We observed 100% cooperation in the experimental sessions of prisoner's dilemma game with the mate choice mechanism, and 7.9% cooperation in the session of the game without the mechanism. Further experimental results and questionnaire analysis find that subjects' behavior is consistent with *BEWDS* rather than *NE*, *SPE* or *NSS* behavior.

JEL Classification Numbers: C72, C73, C92, D74, P43

† This research was supported by the Suntory Foundation, the Joint Usage/Research Center at ISER, Osaka University and "Experimental Social Sciences: Toward Experimentally-based New Social Sciences for the 21st Century" that is a project called the Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science and Culture of Japan. We thank the Economics Department at Osaka University who kindly allowed us to use the computer lab. We also thank comments from participants of seminars at Okinawa, UCLA, Hokkaido, Kyoto Sangyo, UCSB, UCSD, Tohoku, Waseda, Keio, VCASI, Seoul National and 2011 Japanese Economic Association Meeting at Tsukuba. We thank Jiro Akita, Jim Andreoni, Masahiko Aoki, Kenemi Ban, Ted Bergstrom, Tim Cason, Gary Charness, Youngsub Chun, Takako Fujiwara-Greve, Yukihiro Funaki, Yoichi Hizen, Eiji Hosoda, Tatsuya Kameda, Michihiro Kandori, Kazunari Kainou, Shunsuke Managi, Takehito Masuda, Yuko Morimoto, Mayuko Nakamaru, Yusuke Narita, Ichiro Obara, Cheng-Zhong Qin, Junyi Shen, Kazumi Shimizu, Hideo Shinagawa, Martin Shubik, Masanori Takaoka, Toshio Yamagishi, Takehiko Yamato and Bill Zame for their valuable comments.

October 14, 2011
Not for circulation!

We either approve or disapprove of the conduct of another man according as we feel that, when we bring his case home to ourselves, we either can or cannot entirely sympathize with the sentiments and motives which directed it.

Adam Smith, *The Theory of Moral Sentiments*, 1759

1. Introduction

Dresher and Flood conducted the first experiment on Prisoner's Dilemma game at RAND in January of 1950¹, and after their work, numerous papers on its theory and experiments have been published in not only economics but also many fields such as mathematics, computer science, biology, psychology, sociology, political science, management science and so on². There are at least three approaches in order to tame the dilemma.

The first possible direction is to introduce the repetition of the game. David M. Kreps, Paul Milgrom, John Roberts, and Robert Wilson (1982) investigated possible cooperation in *finitely* repeated prisoner's dilemma game. The source of cooperation was some asymmetries of types of players. James Andreoni and John H. Miller (1993) conducted a series of prisoner's dilemma experiments to confirm the prediction by Kreps et al. (1982), and found that subjects' beliefs of their opponent altruism increased reputation building and therefore they were more cooperative than subjects in a repeated single-shot game³. However, the average cooperation scarcely exceeded more than 60%. Yoella Bereby-Meyer and Alvin E. Roth (2006) reported that noisy payoffs reduced cooperation in repeated game although they increased cooperation in one-shot game.

The second approach is related with biology and ecology. Genetic relationship between participants changes the payoff matrix structure called kin selection due to W.D. Hamilton (1964). This idea has been extended to direct, indirect or network reciprocity and group selection (see Michael Doebeli and Christoph Hauert (2005) and Martin A. Nowak (2006) for the review).⁴

¹ According to William Poundstone (1992), "the prisoner's dilemma was "discovered" in 1950, just as nuclear proliferation and arms races became serious concerns" (page 9). See also Merrill M. Flood (1958) and Chapter 6 of Poundstone (1992). Of course, Dresher and Flood were not the first to notice this dilemma problem. David Hume (1739), for example, noticed sequential prisoner's dilemma: "Your corn is ripe to-day; mine will be so to-morrow. 'Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains upon your account; and shou'd I labour with you upon my own account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security." in Book 3.2.5.

² See Alvin E. Roth (1995) for an overview of the experiments.

³ Simon Gächter and Christian Thöni (2005) confirmed that knowing other subjects who are cooperative made subjects cooperative in a public good provision experiment.

⁴ One of their modeling tools is evolutionary dynamics. For example, Christoph Hauert, Arne Traulsen, Hannelore Brandt, Martin A. Nowak, and Karl Sigmund (2007) found how altruistic punishment evolved in the

October 14, 2011
Not for circulation!

Economists such as Robert Sugden (1984), Matthew Rabin (1993) and the followers also have been pursuing this avenue. Rachel T.A. Croson (2007) found that reciprocity plays a key role in linear public good experiments compared with commitment and altruism although it is not good enough to attain the Pareto efficient allocation.

The third approach is to introduce one more stage to the dilemma game in order to *implement* the cooperative outcome.⁵ James Andreoni and Hal Varian (1999) and Gary Charness, Guillaume R. Fréchet and Cheng-Zhong Qin (2007) set up a stage where subjects can reward the other subject conditional upon cooperation before the prisoner's dilemma game stage, called the compensation mechanism. The cooperation rate was about 40-70% in this design. Jeffrey S. Banks, Charles R. Plott and David P. Porter (1988) introduced a voting stage after a public good provision stage as Shubik (2010) suggested, and observed that unanimity reduced efficiency. Although costly punishment does not implement cooperation in a traditionally rational model, following Toshio Yamagishi (1986), Ernst Fehr and Simon Gächter (2000) introduced it in a public good provision experiment, and observed that the average contribution rate was 57.5% with the punishment and 18.5% without it under the stranger matching.⁶

Our approach belongs to the third one. As Elinor Ostrom (1990) showed, many successful examples of the commons usually have some devices before and/or after the strategic decisions of obtaining benefits from the commons, and leaving the dilemma in the commons alone without introducing any devices is extremely rare.⁷ Therefore, our goal is to find a "minimum" reasonable device or mechanism to make players cooperate *theoretically* and *experimentally* in environments as stark as possible. For this end, we assume the behavioral principle that appeared in Hume's quotation in Footnote 1, i.e., all players are absolutely selfish. In addition to that, our challenge is to design a mechanism that is compatible with *reciprocal norm* following the finding by Croson (2007).⁸

model.

⁵ Martin Shubik (2011) emphasized the need a stage *after* prisoner's dilemma game. "Instead of switching to the cooperative game *per se* if the gap were large enough the agents could construct a mechanism in the form of a second stage to the game that provides coordination, signaling and possibly some other forms of control on the original matrix game in such a way that the players can pay for the administrative costs and still all be able to benefit from its existence."

⁶ In addition to this observation, Fehr and Gächter (2000) observed that the average contribution rate was 85% with the punishment and 37.5% without it under the *partner* matching.

⁷ Broadly speaking, our approach is one of "*the Game 5 Ways*" proposed by Ostrom (1990) who set up a contract stage before the dilemma stage called Game 5. This game based upon *empirical* findings is quite different from traditional games that utilize central authority or private property rights.

⁸ For example, the incentive of a costly punisher with a norm ("if you do not cooperate, I will punish you") is the opposite to the incentive of a payoff maximizer. That is, under costly punishment with two players where one is a costly punisher and the other is a payoff maximizer, the punisher chooses defection and the maximizer chooses cooperation.

In order to accomplish this task, we restrict ourselves to the class of mechanisms satisfying the following two stringent conditions. First, they must be "*onto*". That is, the four possible outcomes of prisoner's dilemma game are exactly the same as the outcomes of the mechanisms, and hence the outcomes other than these four should not be used. This implies that they do not accept any payoff flow from or to the outside. In this sense, they must be budget-balanced.⁹ For example, a mechanism that gives some monetary payoff to a player who chooses cooperation from the outside is not onto. Furthermore, we impose not using *direct* punishment (or reward) since *personal* punishment (or bribe) is usually prohibited in our modern societies or legal systems.¹⁰ Second, the mechanisms must be "*voluntary*". Any player who chooses defection should not be forced to change the decision to cooperation.

Under the above constraints, we introduce the *approval* stage after the prisoner's dilemma as Adam Smith (1759) suggested. After the dilemma stage, each subject can approve ("*yes*") or disapprove ("*no*") the other choice of the strategy in the first stage. Although there are many ways to design the rule or the mechanism, we employ the following simple one called the *mate choice mechanism*: if both approve the other strategy, the outcome is the one with which both choose in the first stage, and if either one disapproves it, the outcome is the one with which both defect in the first stage¹¹. Apparently, the mate choice mechanism satisfies the onto and voluntary conditions. Furthermore, this mechanism satisfies *forthrightness* saying that the outcome must be what players choose whenever both approve the other choices. We also show that the mate choice mechanism is *unique* satisfying forthrightness and several other conditions.

Our basic behavioral principle is that each player is a payoff maximizer. We also consider a *reciprocator*: a reciprocator chooses "*yes*" if the other player chooses to cooperate, and the player chooses "*no*" if not.¹² We assume that a reciprocator is a payoff maximizer. That is, a reciprocator has a norm that regulates the behavior, but (s)he maximizes the payoff as far as (s)he follows the norm. Therefore, our research question is whether or not the mate choice mechanism can align incentives of players who are payoff maximizers and/or reciprocators theoretically and experimentally. Of course, the behavior depends upon an equilibrium concept employed.

We prepare five possible equilibrium concepts: Nash equilibrium (*NE*), subgame perfect

⁹ This is a standard condition in mechanism design. See, for example, Chapter 23 of Andreu Mas-Colell, Michael D. Whinston and Jerry R. Green (1995).

¹⁰ Costly punishment does not satisfy the onto condition. Francesco Guala (2010) surveyed literature including ethnology, anthropology and biology, and concluded that costly punishment was rare.

¹¹ Although Raúl López-Pérez and Marc Vorsatz (2010) also investigated the approval stage after the prisoner's dilemma game, their design at the stage did not affect the final outcomes, and the cooperation rates were 22-38%.

¹² The typical definition of reciprocity is "if the other cooperate, then I will cooperate" (see, for example, Robert Sugden (1984), Matthew Rabin (1993) and Rachel Croson (2007) among others). Since our game has two stages, we take advantage of this structure, and regard approval as a norm of reciprocity.

equilibrium (*SPE*), evolutionarily stable strategies (*ESS*), neutrally stable strategies (*NSS*), and backward elimination of weakly dominated strategies (*BEWDS*).¹³ If both are payoff maximizers, then the *NE* and *SPE* have all combinations of cooperation (*C*) and defection (*D*), but no *ESS* exists. On the other hand, the outcome of *NSS* and *BEWDS* is that both choose cooperation. If both are reciprocators, then the *NE* and *SPE* have (*D,D*) as an outcome, and the outcome of *NSS*, *ESS*, and *BEWDS* is (*C,C*). If one player is a payoff maximizer and the other is a reciprocator, then the *NE* and *SPE* have (*D,D*) as an outcome, and the outcome of *BEWDS* is (*C,C*). In this sense, the mate choice mechanism does not implement the cooperative outcome in either *NE* or *SPE*, but it implements the outcome in *NSS* if both are payoff maximizers or both are reciprocators, in *ESS* if both are reciprocators and in *BEWDS* without specifying any restriction.

Therefore, our experimental task is to find how subjects cooperate and to identify which equilibrium concepts they choose. In our experimental design, we aim at constructing the environment as bleak as possible against cooperation. In order to avoid possible learning or building-up reputation, no subject ever met another subject more than once, called the complete stranger design.¹⁴ Furthermore, each subject could not identify where the other subject was located in the lab. As usual in this type of experiment, no talking was allowed.

Our observation is rather striking. We observe 100% cooperation in the session of prisoner's dilemma game with the mate choice mechanism in 19 periods, and 7.9% cooperation in the session of the game without the mechanism.

We also check the robustness of the mate choice mechanism with two additional and slightly different sessions. The first one is prisoner's dilemma game with *unanimous voting* where we change the wording for the mate choice mechanism, but keep the game-theoretical structure. After prisoner's dilemma game decision, each subject votes "yes" or "no" to the choice pair in the first stage. Whenever both choose "yes", the choice pair is finalized. If either one says "no", the outcome when both choose defection is selected. This mechanism is mathematically equivalent to the mate choice mechanism and we observe that the average cooperation rate is 98.2%.

The second one is the prisoner's dilemma game with the mate choice mechanism *without* repetition. All subjects of ten pairs chose cooperation in the first stage, and then chose approval in the second stage. These three sessions show that the mate choice mechanism is robust enough to attain the full or almost full cooperation.

¹³ R. Selten (1975) is the initiator who used the idea of *BEWDS* in game theory, and later Ehud Kalai (1981) used *BEWDS* in the *PD* Game and Banks, Plott and Porter (1988) used it in the provision of a public good in implementing cooperation.

¹⁴ John Duffy and Jack Ochs (2009) reported that random matching treatment in a repeated prisoner's dilemma game failed to generate cooperative norm contrary to a theoretical prediction by Michihiro Kandori (1992).

Which equilibrium concept is compatible with the data? The set of equilibria based upon *BEWDS* is a proper subset of the set of them based upon *NSS*. The latter is also a proper subset of the set of them based upon *NE* or *SPE*. Almost all data fall into the set of equilibria based upon *BEWDS*. In this sense, *BEWDS* is the most compatible among the equilibria.

The mate choice mechanism reduces cognitive burden of subjects under *BEWDS*. Subjects who use *BEWDS* must compare two dimensional vectors at each subgame after the choice of either cooperation or defection in the prisoner's dilemma game stage. Notice that a payoff vector (u,v) weakly dominates (x,y) if $u \geq x$ and $v \geq y$ and at least one strict inequality. If either one disapproves the other choice in the second stage, then all three payoff vectors out of four at the subgame are the same, that is called the *mate choice flat*. Therefore, subjects must compare just two numbers a and c , not two vectors since $b = d$ due to the flat. Furthermore, this made subjects think backwardly easier than the situation without the flat. That is, we have strong evidence where subjects considered the two stage game backwardly together with the eliminations.

Our finding of *BEWDS* as the behavioral principle is in a limited environment, however, it opens up a new avenue to design mechanisms not using *NE*, *SPE*, *ESS* or *NSS*, but using subjects' ability to use weak dominance of strategies and to think a stage game backwardly.

In order to compare the above results, we used the compensation mechanism by Andreoni and Varian (1999) before the prisoner's dilemma game that has the same symmetric payoff table in the above experiments although they used an asymmetric payoff table.¹⁵ The outcome when both cooperate is the unique *SPE* in the prisoner's dilemma game with the compensation mechanism although all possible combinations of *C* and *D* are the outcomes of *BEWDS* assuming that both are payoff maximizers. Our finding with 19 rounds was that the average cooperation rate is 76.6% that is higher than that in their experiment, but it is significantly different from the rate of the mate choice mechanism.

Although it is notorious, a good example of the mate choice mechanism 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.¹⁶ 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

¹⁵ See also Hal R. Varian (1994).

¹⁶ We thank Toshio Yamagishi who pointed out this example.

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.¹⁷ 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 Bruce Russett, Harvey Starr and David 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.¹⁸ 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. Adding the second stage in resolving conflicts has been used widely in our societies.

The organization of the paper is as follows. Section 2 explains the mate choice mechanism as a special case of the approval mechanism. Sections 3, 4, 5 and 6 are for theoretical properties of the mate choice mechanism under *NE*, *SPE*, *ESS*, *NSS*, *BEWDS* and reciprocity. Section 7 takes care of various possibilities of implementation. Section 8 describes experimental procedures and section 9 is for experimental results. Section 10 explains why *BEWDS* works well and then characterizes the mate choice mechanism. Section 11 is for further research agenda.

2. Prisoner's Dilemma with Approval Mechanism

¹⁷ 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.

¹⁸ We thank Kazunari Kainou who provided most of the examples in this paragraph.

The prisoner's dilemma (PD) game with approval mechanism consists of two stages. In the first stage, players 1 and 2 face a usual PD game such as Figure 1. In each cell, the first number is the payoff for player 1 and the second is for player 2. Both players must choose either cooperation (C) or defection (D) simultaneously. There might be many ways to interpret the matrix in Figure 1, but a typical interpretation in public economics is the payoff matrix of the voluntary contribution mechanism in the provision of a public good. Each player has ten dollars (or initial endowment w) at the beginning, and (s)he must decide whether (s)he contributes all ten dollars (or cooperates) or nothing (or defects). The sum of the contribution is multiplied by $\alpha \in (0.5, 1)$, that is 0.7 in the following example, and the benefit goes to both of them, which expresses non-rivalness of the public good. If both contribute, then the benefit of each player is $(10+10) \times 0.7 = 14$. If either one of them contributes, contributor's benefit is $10 \times 0.7 = 7$, and non-contributor's benefit is $10 + 7 = 17$ since (s)he has 10 dollars at hand. Therefore, the payoff matrix in Figure 1 keeps this linear structure. Of course, non-contribution (D) is the dominant strategy.¹⁹ Bold and italic numbers in the lower right cell show the equilibrium payoff in Figure 1.

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

Figure 1. Prisoner's dilemma game.

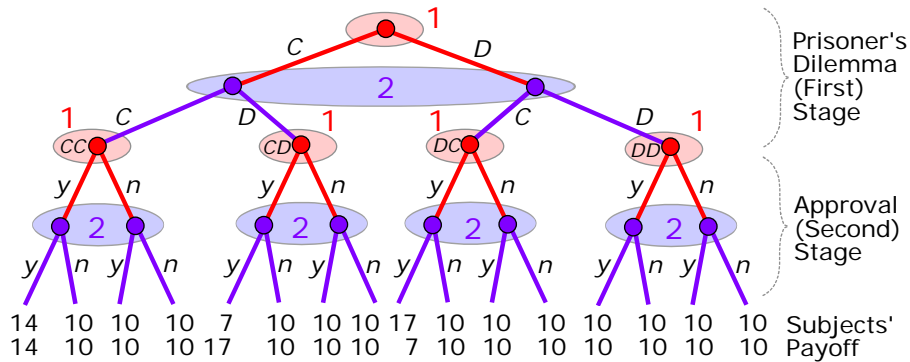


Figure 2. Prisoner's dilemma game with the mate choice mechanism.

Consider now the second stage as in Figure 2. Knowing the strategy pair of the first stage, each player must either approve the strategy choice of the other (y) or disapprove it (n) simultaneously. Ellipses show the information sets. Since each set has two alternatives, and there

¹⁹ In the experiment, we used payoff numbers that are 100 times of the numbers in Figures 1 and 2 due to the exchange rate.

are ten information sets, the total number of possible strategy profiles is $1024 (= 2^{10})$. The upper (lower) number at the bottom of the game tree show player 1's (2's) payoff respectively.

Although there are many ways to connect the approval decisions to the strategy choices in the first stage, we choose the following simple way since this procedure has a special feature on the uniqueness of the approval mechanism that will be discussed later: 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) that corresponds to (D,D) in the first stage. In the context of public good provision, when either one of them disapproves the choice of the other, the public good will not be provided and hence the money is simply return to the contributors.²⁰

Another interpretation of the above specification of the approval mechanism is *mate choice*. A male and a female meet together. If both approve the other, then they can make a mate. If either one of them disapproves the other choice, then they must stay at the status quo since they cannot be a mate. We call this specification of approval mechanism the *mate choice mechanism* (*MCM*).²¹ The *PD* game with *MCM* in Figure 2 is abbreviated as *PDMC*.²²

We will consider several equilibrium concepts whose equilibrium outcomes are different for the next four sections. Sections 3, 4, and 5 are for payoff maximizers, and section 6 is for reciprocators or the mixture of a payoff maximizer and a reciprocator.

3. Nash and Subgame Perfect Equilibria of the *PDMC*

An equilibrium concept that has been widely used in analyzing the two stage game is subgame perfect equilibrium (*SPE*). Consider four subgames in the second stage in Figures 2 and 3. Bold and italic numbers in a cell show that the pair is Nash equilibrium (*NE*) and some of them are not black but gray that are eliminated under *BEWDS* in the next section. Subgame *CC* has two *NEs* ((y,y) and (n,n)). Similarly, subgame *CD* has two *NEs* ((n,y) and (n,n)), subgame *DC* has two *NEs* ((y,n) and (n,n)), and subgame *DD* has four *NEs* ((y,y) , (y,n) , (n,y) and (n,n)).

The four subgames have a point in common: the payoff at (y,n) , (n,y) and (n,n) is (10,10) and hence it is *flat*, that we call the *mate choice flat*, in the matrices due to the *MCM*. Any payoff that is lower than the status quo payoff, i.e., 10, would never be an *NE*. That is, $(7,17)$ or $(17,7)$

²⁰ This specification is different from so called the money back guarantee mechanism. Consider the mechanism if either one of the two chooses *C* but not both, then the 10 contribution is returned to the cooperators. This mechanism cannot generate $(7,17)$ where (C,D) in the first stage and both choose *y* in the second stage in Figure 2.

²¹ There are six approval rules whose outcome is exactly the same as the mate choice rule. See Tatsuyoshi Saijo and Yoshitaka Okano (2009).

²² The definition of mate choice in biology is much broader than our usage: according to T. R. Halliday (1983), "Mate choice may be operationally defined as any pattern of behaviour, shown by members of one sex, that leads to their being more likely to mate with certain members of the opposite sex than with others."

would not be chosen and the mechanism prevents free-riding. In this sense, the mechanism is a device for *survival* not to end up at payoff 7. Another common point is that (n,n) is always an *NE* due to the mate choice flat. This makes Pareto inferior payoff vector $(10,10)$ to $(14,14)$ survive as an equilibrium, and hence this needs an equilibrium refinement or different equilibrium concepts to exclude $(10,10)$.

		Player 2				Player 2				Player 2	
		y	n	y	n	y	n	y	n	y	n
Player 1	y	14,14	10,10	7,17	10,10	17,7	10,10	10,10	10,10	10,10	10,10
	n	10,10	10,10	10,10	10,10	10,10	10,10	10,10	10,10	10,10	10,10
		Subgame CC		Subgame CD		Subgame DC		Subgame DC		Subgame DD	

Figure 3. Four subgames in *PDMC*.

Given the outcomes of four subgames, we can construct the reduced normal form games. Since the payoff of all *NEs* in subgames *CD*, *DC* and *DD* is $(10,10)$, consider two cases (y,y) and (n,n) in subgame *CC*. If it is (y,y) in subgame *CC*, there are two *NEs* (C,C) and (D,D) in the reduced normal form game (see Figure 4-(i)). Since each equilibrium has 16 cases (i.e., two *NEs* in subgame *CD*, two in subgame *DC*, and four in subgame *DD*), there are 32 *SPEs*. On the other hand, if it is (n,n) in subgame *CC*, there are four *NEs* (C,C) , (C,D) , (D,C) and (D,D) (see Figure 4-(ii)). Since each equilibrium has 16 cases, we have 64 *SPEs*. In total, there are 96 *SPEs*. Notice also that two games in Figure 4 have the mate choice flat.

		Player 2				Player 2	
		C	D			C	D
Player 1	C	14,14	10,10	Player 1	C	10,10	10,10
	D	10,10	10,10	D	D	10,10	10,10
(i) (y, y) in subgame CC				(ii) (n, n) in subgame CC			

Figure 4. The prisoner's dilemma stage in the backward induction.

Consider the *SPE* paths in Figure 2.²³ First, fix (y,y) at subgame *CC*. Then they are (C,C,y,y) (16 cases), (D,D,y,y) (4 cases), (D,D,y,n) (4 cases), (D,D,n,y) (4 cases), and (D,D,n,n) (4 cases). Second, fix (n,n) at subgame *CC*. Then they are (C,C,n,n) (16 cases), (C,D,n,y) (8 cases), (C,D,n,n) (8 cases), (D,C,y,n) (8 cases), (D,C,n,n) (8 cases), (D,D,y,y) (4 cases), (D,D,y,n) (4 cases),

²³ A *path* is (subject 1's choice between C and D, subject 2' choice between C and D, subject 1's choice between y and n, subject 2's choice between y and n).

(D,D,n,y) (4 cases), and (D,D,n,n) (4 cases).

Consider *NEs* of the game in Figure 2. Since each player has 32 strategies, we can construct a 32 by 32 payoff matrix. Finding the intersection of best responses of two players, we have 416 *NEs* with equilibrium paths of (C,C,y,y) (16 cases), (D,D,y,y) (64 cases), (D,D,y,n) (64 cases), (D,D,n,y) (64 cases), (D,D,n,n) (64 cases), (C,C,n,n) (16 cases), (C,D,n,y) (32 cases), (C,D,n,n) (32 cases), (D,C,y,n) (32 cases), and (D,C,n,n) (32 cases). Summarizing these, we have,

Property 1. *In the PDMC, we have*

- (i) 416 *NEs* and 96 *SPEs* out of 1024 possible strategy profiles;
- (ii) the *NE* paths and the *SPE* paths are the same and they are (C,C,y,y) , (D,D,y,y) , (D,D,y,n) , (D,D,n,y) , (D,D,n,n) , (C,C,n,n) , (C,D,n,y) , (C,D,n,n) , (D,C,y,n) , and (D,C,n,n) ; and
- (iii) the payoff of 16 *NEs* and 16 *SPEs* is $(14,14)$ on the path (C,C,y,y) and the payoff of the rest is $(10,10)$.

Let us define player i 's strategy of the two stage game as $s_i = (E_i, s_i^{CC}, s_i^{CD}, s_i^{DC}, s_i^{DD})$ where E_i is i 's choice between C and D in the *PD* stage, and s_i^{AB} is i 's choice between y and n in the *MCM* when player i chooses A and player j chooses B in the *PD* stage.²⁴ Then we have,

Property 2. 16 strategy profiles where the outcome of *SPE* is (C,C) are

$(s_1, s_2) = ((C, y, n, \cdot, \cdot), (C, y, n, \cdot, \cdot))$ where " \cdot " indicates either y or n .

4. Neutrally Stable Strategies of PDMC

This section presents neutrally stable strategies (*NSS*), which is a refinement of *NE*, of the *PDMC*. We will show that all *NSS* paths are (C,C,y,y) , and the game does not have evolutionarily stable strategy. Since players 1 and 2 are symmetric, let us abbreviate player's subscript and let $v(s,t)$ be the payoff of player 1 when the strategy profile is (s,t) . Due to payoff symmetry, player 2's payoff at (s,t) is $v(t,s)$.

Definition 1. A strategy t is a *neutrally stable strategy* if and only if for all $t' \neq t$,

- (i) $v(t,t) \geq v(t',t)$ and
- (ii) $v(t,t) = v(t',t)$ implies $v(t,t') \geq v(t',t')$.

If the weak inequality in (ii) becomes strict, then t is called an *evolutionarily stable strategy* (*ESS*).

²⁴ For example, consider $s_1 = s_2 = (C, y, n, y, y)$. This indicates that player 1 chooses n and player 2 chooses y at subgame *CD*, and player 1 chooses y and player 2 chooses n at subgame *DC*.

Property 3. A strategy t is an NSS if and only if $t = (C, y, n, \cdot, \cdot)$ where " \cdot " indicates either y or n .

Proof. See Appendix.

Combining Properties 2 and 3, we have,

Property 4. All 16 NSS profiles are exactly the same as the 16 SPE profiles whose payoff is (14,14).

5. Backward Elimination of Weakly Dominated Strategies of the PDMC

Another equilibrium concept whose outcome exactly coincides with (C,C) is backward elimination of weakly dominated strategies (BEWDS) which is also adopted, for example, in Ehud Kalai (1981). This requires two properties. The first is subgame perfection and the second is that players do not choose weakly dominated strategies in each subgame and the reduced normal form game.

Let us take a look at the subgame whose starting node is CC in Figure 2. Notice that (14,10) corresponds to y and (10,10) to n for both players. We say strategy α with (u,v) weakly dominates strategy β with (x,y) if $u \geq x$ and $v \geq y$ with at least one strict inequality or $(u,v) \geq (x,y)$. That is, since y weakly dominates n , n should not be chosen. Therefore, (y,y) is realized at subgame CC. Similarly, (n,y) at subgame CD and (y,n) at subgame DC are realized. In subgame DD, since no weakly dominated strategy exists, (y,y) , (y,n) , (n,y) and (n,n) are realized. Given the realized strategies in all subgames, we have the reduced normal form game in Figure 4-(i). In this game, C weakly dominates D for both players and hence, (C,C) is the realized outcome. Since there are four realized pairs in subgame DD, there are four realized predictions in the two stage game. Notice that the order of elimination in each stage does not change the final outcome.

Property 6. Using BEWDS in PDMC, we have

- (i) four realized predictions; and
- (ii) the unique prediction path is (C,C,y,y) with the payoff of (14,14).

6. Approval as a Norm of Reciprocity

Thus far, the payoff maximization is the objective of each player. Following Croson (2007), we impose the following reciprocal norm: if the other chooses C, then I will approve it, and if not, then I will disapprove it. We call a payoff maximizer who has this norm a *reciprocator* (R). Maximizing behavior depends on equilibrium concepts. There are two cases: both are

reciprocators, and one of them is a reciprocator and the other is payoff maximizer.

Let us consider the case where both are reciprocators. Then the reciprocal norm implies $s_i = (E_i, s_i^{CC}, s_i^{CD}, s_i^{DC}, s_i^{DD}) = (\cdot, y, n, y, n)$. The payoff maximization behavior and equilibrium concepts employed determine the first element of s_i . Suppose first that both are *R*s. Then, the reduced normal form game is exactly the same as the payoff matrix of Figure 4-(i). Therefore, the choices of *NE* and *SPE* are (C,C) and (D,D) . By definition, the choice of *NSS*, *ESS*, and *BEWDS* is (C,C) .

		Player 2																											
		<i>y</i>	<i>n</i>	<i>y</i>	<i>N</i>																								
Player 1	<i>y</i>	<table border="1" style="display: inline-table; border-collapse: collapse;"><tr><td><i>y</i></td><td>14,14</td><td>10,10</td></tr><tr><td><i>n</i></td><td>·,10</td><td>·,10</td></tr></table>	<i>y</i>	14,14	10,10	<i>n</i>	·,10	·,10	<table border="1" style="display: inline-table; border-collapse: collapse;"><tr><td><i>y</i></td><td>·,17</td><td>·,10</td></tr><tr><td><i>n</i></td><td>10,10</td><td>10,10</td></tr></table>	<i>y</i>	·,17	·,10	<i>n</i>	10,10	10,10	<table border="1" style="display: inline-table; border-collapse: collapse;"><tr><td><i>y</i></td><td>17,7</td><td>10,10</td></tr><tr><td><i>n</i></td><td>·,10</td><td>·,10</td></tr></table>	<i>y</i>	17,7	10,10	<i>n</i>	·,10	·,10	<table border="1" style="display: inline-table; border-collapse: collapse;"><tr><td><i>y</i></td><td>·,10</td><td>·,10</td></tr><tr><td><i>n</i></td><td>10,10</td><td>10,10</td></tr></table>	<i>y</i>	·,10	·,10	<i>n</i>	10,10	10,10
	<i>y</i>	14,14	10,10																										
<i>n</i>	·,10	·,10																											
<i>y</i>	·,17	·,10																											
<i>n</i>	10,10	10,10																											
<i>y</i>	17,7	10,10																											
<i>n</i>	·,10	·,10																											
<i>y</i>	·,10	·,10																											
<i>n</i>	10,10	10,10																											
		Subgame <i>CC</i>	Subgame <i>CD</i>	Subgame <i>DC</i>	Subgame <i>DD</i>																								

Figure 5. Subgames when player 1 is a reciprocator and player 2 a payoff maximizer.

Consider the case where player 1 is an *R* and player 2 is a payoff maximizer. Then strategies for player 1 are *C* and *D*, i.e., $E_1 = C$ or D since the choices of subgames are determined by the norm, i.e., $(s_1^{CC}, s_1^{CD}, s_1^{DC}, s_1^{DD}) = (y, n, y, n)$, and player 2 has $2^5 = 32$ strategies. Then there are 24 *NE*s with equilibrium paths of (C,C,y,y) , (D,C,y,n) , (D,D,n,y) and (D,D,n,n) . The cells with bold square in Figure 5 show the outcomes of each subgame, and hence we can obtain the matrix of Figure 4-(i) and 8 *SPE*s with equilibrium paths of (C,C,y,y) , (D,D,n,y) and (D,D,n,n) . Consider *BEWDS*. Since player 2 chooses *y* at subgame *CD* in Figure 5, and *y* and *n* are indifferent in subgame *DD*, $s_2 = (E_2, s_2^{CC}, s_2^{CD}, s_2^{DC}, s_2^{DD}) = (C, y, n, y, \cdot)$. That is, we have two realized predictions in the normal form game, and the unique prediction path is (C,C,y,y) . The following property summarizes the results.²⁵

Property 6. *In PDMC, we have*

- (i) *if both players are reciprocators, the choices of NE and SPE are (C,C) and (D,D), and the choices of NSS, ESS and BEWDS is (C,C);*
- (ii) *if player 1 is a reciprocator and player 2 is a payoff maximizer,*
 - (ii-1) *there are 24 NEs with the equilibrium paths of (C,C,y,y), (D,C,y,n), (D,D,n,y) and (D,D,n,n), and 8 SPEs with the equilibrium paths of (C,C,y,y), (D,D,n,y) and (D,D,n,n); and*
 - (ii-2) *the unique prediction path of BEWDS is (C,C,y,y).*

7. Implementability

²⁵ Since the game is asymmetric, we did not find *ESS* or *NSS*.

We will consider implementability of the MCM in an economic environment with a public good.²⁶ Let $u_i(x_i, y) = x_i + \alpha_i y$ be a utility function defined on R_+^2 where x_i is a private good, y is a public good, and let $\mathbf{U} = \{(u_1, u_2) : u_i = x_i + \alpha_i y \text{ for some } \alpha_i \in (0, 5, 1)\}$. Let $y = h(x) = \sum t_i$ be a production function of the public good where $t_i = w_i - x_i$ and w_i is player i 's initial endowment. Then let $A = \{(x_1, y), (x_2, y)\} \in R_+^4 : y = \sum (w_i - x_i)\}$ be the set of feasible allocations. Define a social choice correspondence $f : \mathbf{U} \rightarrow A$ by $f(u) =$ the set of maximizers of $\sum u_i(x_i, y)$ on A . Apparently, this correspondence is a function in our setting, and the optimal level of the public good is $w_1 + w_2$. Let $g : \mathbf{S} \rightarrow A$ be a game form (or mechanism) where \mathbf{S} is the set of strategy profiles, and let $E_g : \mathbf{U} \rightarrow \mathbf{S}$ be the equilibrium correspondence based upon equilibrium concept E . Then we say that mechanism g implements f in E if $f(u) = g \cdot E_g(u)$ for all u . In our special case, we set $w_1 = w_2 = 10$ and $\alpha_1 = \alpha_2 = 0.7$. We also regard $t_1 = t_2 = 10$ as C , and $t_1 = t_2 = 0$ as D . Furthermore, we do not allow any number between 0 and 10.²⁷

Consider first that both are payoff maximizers. In the *PD* case, the strategy space for each player is $\{0, 10\}$ and the game form is $h(t_1, t_2) = ((x_1, y), (x_2, y))$ with $y = \sum t_i$. Let D_h be the set of dominant strategy equilibria. Then since $h \cdot D_h(u) = ((w_1, 0), (w_2, 0))$, h cannot implement f in dominant strategy equilibria.

Let g be the MCM. Write $BEWDS_g(u)$ as the set of realized predictions by BEWDS using g under u . Then $g \cdot BEWDS_g(u) = f(u)$ and hence g implements f in BEWDS. Similarly, writing NSS_g as the set of NSS pairs, we have $g \cdot NSS_g(u) = f(u)$ for all possible u . That is, g implements f in NSS²⁸. Let $NE_g(u)$ ($SPE_g(u)$) be the set of NEs (*SPEs*) using g under u . Then $g \cdot NE_g(u) \neq f(u)$ ($\neq g \cdot SPE_g(u)$), and hence g cannot implement f in NE (or *SPE*). Obviously, g cannot implement f in ESS since no ESS exists. Apply the same procedures to the cases in section 6, we have,

Property 7. (i) Suppose that both players are payoff maximizers or reciprocators, or the mixture of them.

Then,

(i-1) Cooperation cannot be attained in the *PD* game in dominant strategy;

(i-2) The MCM cannot implement cooperation of *PD* in either NE or *SPE*; and

²⁶ See Eric Maskin (1999) and Saijo (1988) for Nash implementation, John Moore and Rafael Repullo (1988) for *SPE* implementation, and Matthew O. Jackson (2001) for a general survey.

²⁷ As Takehito Masuda, Okano and Saijo (2011) shows, if the number of strategies is more than two, the MCM fails to implement the social choice correspondence in BEWDS. That is, there is a deep chasm between two and three strategies.

²⁸ Mayuko Nakamaru, Saijo and Takehiko Yamato (2011), as a companion paper to the current paper, show that the prisoner's dilemma game with approval stage implements the Pareto optimal outcome in an evolutionary dynamics model.

- (i-3) The MCM implements cooperation of PD in BEWDS.
 (ii) Suppose that either both players are payoff maximizers or reciprocators. Then,
 (ii-1) the MCM implements cooperation of PD in NSS; and
 (ii-2) the MCM cannot implement cooperation of PD in ESS if both players are payoff maximizers, but it implements cooperation of PD in ESS if both are reciprocators.

	Both are Payoff Maximizers (sections 3,4 & 5)	Both are Reciprocators (section 6)	Mixture of the Two (section 6)
<i>Nash</i>	(C,C), (C,D),(D,C),(D,D)	(C,C), (D,D)	(C,C), (C,D),(D,C),(D,D)
<i>SPE</i>	(C,C), (C,D),(D,C),(D,D)	(C,C), (D,D)	(C,C), (D,D)
<i>NSS</i>	(C,C)	(C,C)	N.A.
<i>ESS</i>	No ESS	(C,C)	N.A.
<i>BEWDS</i>	(C,C)	(C,C)	(C,C)

N.A.: not applicable due to asymmetry of the game.

Table 1. Outcomes of PDMC under Five Equilibrium Concepts.

The MCM implements cooperation with two types of players including the mixture of them under BEWDS. That is, the mechanism implements a social goal under *an* equilibrium concept even there are *two* behavioral principles among players. In this sense, we name this *bipartite* implementation. Notice that this is different from *double* implementation in the literature. Given one behavioral principle such as payoff maximizing behavior, a mechanism implements some social goal with *two* equilibria in double implementation.²⁹ Table 1 summarizes the results in sections 3-7 regarding the outcomes of PDMC under five equilibrium concepts. The dark areas show that (C,C) is the unique outcome and hence cooperation is implementable.

8. Experimental Procedures

We conducted the experiment for a day in November 2009, a day in March 2010 and two days in November 2010 at the Economics Department Computer Laboratory at Toyonaka campus of Osaka University. We had two sessions in November 2009: one of them was the PD session, and the other was the PDMC session. We also had two sessions in March 2010 named the PD with unanimous voting (PDUV) and the PD with compensation mechanism (CMPD) sessions explained later. Each of these sessions had 19 rounds. Two sessions in November 2010 were PD* and PDMC* where "*" showed no repetition.

²⁹ For the recent development of double implementation and its experiment, see Saijo, Tomas Sjöström, and Yamato (2007) and Timothy Cason, Saijo, Sjöström and Yamato (2006). As for bipartite implementation, the MCM is the first mechanism that explicitly deals with multiple behavioral principles under one equilibrium concept to the best of our knowledge.

Twenty subjects participated in each session, and hence the total number of subjects was 120. No subjects participated in more than one session. We recruited these subjects by campus-wide advertisement where subjects' information was summarized in Table A1 in the appendix. They were told that there would be an opportunity to earn money in a research experiment. Communication among the subjects was prohibited, and we declared that the experiment would be stopped if it was observed. This never happened. The experiment required approximately 75 minutes to complete in the *PD* session, 115 minutes in the *PDMC* session, 98 minutes in the *PDUV* session, 132 minutes in the *CMPD* session, 72 minutes in the *PD** a session and 80 minutes in the *PDMC** session.

The experimental procedure is as follows. We made ten pairs out of twenty subjects seated at computer terminals in each session³⁰. The pairings were anonymous and were determined in advance so as not to pair the same two subjects more than once in sessions with repetitions. Since most of the previous studies such as Andreoni and Varian (1999) (Charness, Fréchet and Qin (2007)) employed random matching among 4 to 8 subjects (2 to 4 groups)³¹, the repetition necessarily entails of pairings of the same two subjects. Therefore, compared to the previous experiments, this “complete” strangers design might reduce possibility of cooperation among subjects.³² Each subject received instruction sheet and record sheet. The instruction was read loudly by the same experimenter.

Let us explain the *PDMC* session. Before the real periods started, we allowed the subjects five minutes to examine the payoff table and to consider their strategies. 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 filled in it on the record sheet. After that, each subject wrote the choice reason in a small box on the record sheet by hand. Then the 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 inputted the decision into a computer and also filled in it on the record sheet. After that, each subject wrote the reason in a small box by hand. Once every subject finished the task, each subject 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 session without the decision stage became the *PD* session. After finishing all 19 periods, every subject filled in questionnaire

³⁰ We used the z-Tree program developed by Urs Fischbacher (2007).

³¹ Charness et al. (2007) partitioned 16 subjects in one session into four separate groups, with the 4 subjects in each group interacting only with each other over the course of the session.

³² An exception is Cooper et al. (1996) who employed the complete stranger matching.

sheets. The $PDMC^*$ and PD^* sessions were exactly the same as the $PDMC$ and PD sessions without repetition, respectively.

In order to examine the robustness of the mate choice mechanism and to understand the framing effect, we also conducted the PD game with unanimous voting ($PDUV$) session. The experimental procedure is exactly the same as in the $PDMC$ session except for the unanimous voting stage. Each subject must vote for the outcome of the PD stage. If both affirm the strategy choices in the PD stage, then the outcome is what they choose in the PD stage. Otherwise, the outcome is (10,10). That is, $PDMC$ and $PDUV$ are mathematically equivalent, but not *cognitively*. For example, suppose that (C,D) (or (B,A) in the experiment) is observed in the PD stage. In $PDMC$, subject 1 is asked to choose either approve or disapprove subject 2's choice D , but in $PDUV$, subject 1 is asked to vote on the outcome (C,D) . In this sense, comparing $PDMC$ with $PDUV$ is to understand the framing effect.

We also compare our results with two-stage game experiments introduced by Andreoni and Varian (1999). They added a stage called the *compensation mechanism* (CM) where each subject could offer to pay the other subject to cooperate *before* the PD stage. Then they showed that the unique SPE outcome was Pareto efficient in their asymmetric payoff table although all possible combinations of C and D are the outcomes of $BEWDS$ assuming that both are payoff maximizer.³³ Eight subjects in a group formed four groups and the matching was random. They played a usual PD game for the first 15 periods, and then played the two stage game from 16 to 40 periods. The cooperation rate of the former was 25.8% and the latter was 50.5%.

We used the PD game in Figure 1 rather than their asymmetric game. We refer to this session as the $CMPD$ session. The unique SPE outcome is that both offer three in the CM , and choose cooperation in the PD stage. Due to discreteness of strategies, the equilibrium offers are either three or four in the CM . On the other hand, all possible combinations such as (C,C) , (C,D) , (D,C) and (D,D) fall into $BEWDS$ with equilibrium offer of three. When the offer takes discrete value, the $BEWDS$ outcomes are $(3,3,C,C)$, $(3,3,C,D)$, $(3,3,D,C)$, $(3,3,D,D)$, $(3,4,C,C)$, $(3,4,C,D)$, $(4,3,C,C)$, $(4,3,D,C)$ and $(4,4,C,C)$.

We used twenty subjects in each session with repetition. There would have been using, for example, four subjects in a session and repeating five times. We did not take this design in order to avoid the repetitious pairing of the same subjects and to keep the "complete" stranger matching. That is, we made our experimental environment as barren as possible against cooperation. This change does not affect the statistical tests when we compare the cooperation

³³ For further details, see Okano and Saijo and Junyi Shen (2011).

rates across sessions because we have twenty data in both cases³⁴.

9. Experimental Results

9.1. The Effect of the Mate Choice Mechanism

Let us take a look at the *PDMC* session first. All twenty subjects chose cooperative strategy in the choice (or *PD*) stage, and then chose approval in the decision (or approval) stage in all nineteen periods. See also Figure 6.

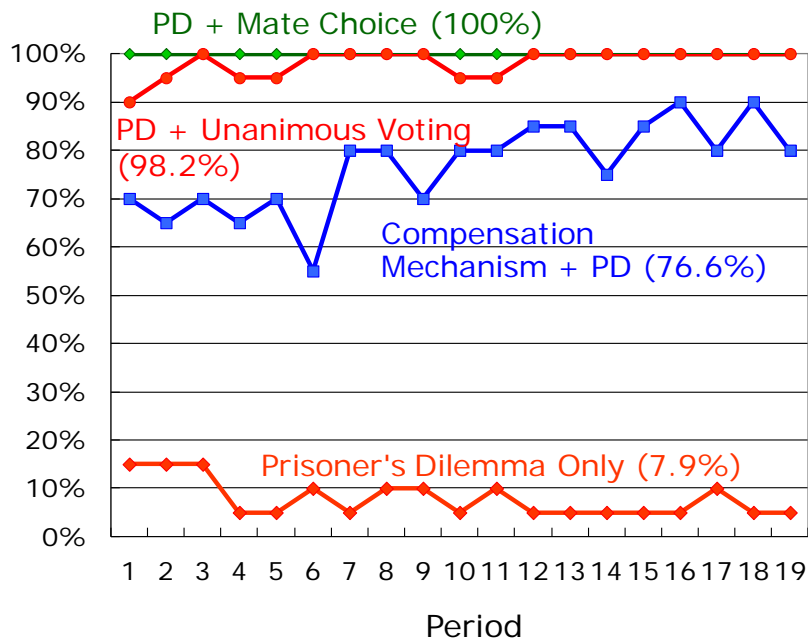


Figure 6. Cooperation rates of four sessions.

In the *PD* session, subject 5 chose cooperation in periods 1, 2 and 3, subject 10 chose it in periods 3 and 8, subject 13 chose it in periods 1 and 11, subject 16 chose it in periods 2, 6, 9 and 17, and subject 17 chose it in all periods. In total, the number of cooperation choice is 30 out of 380, which is 7.9%, and 11% for the first five periods declining to 6% for the last five periods. No (C,C) was observed among 190 pairs of choices. The cooperation rate in our experiment is slightly lower than the previous experiments. For example, Alvin E. Roth and J. Keith Murnighan (1978) find 10.1% cooperation, Russell Cooper et al. (1996) find 20%, and Andreoni and Miller (1993) find 18%. Hence, our subjects are more in line with game-theoretic logic such as adopting dominant or Nash equilibrium strategy. The difference of cooperation rates between *PDMC* and *PD* session is

³⁴ We will borrow and show the data of the other sessions of *PDMC* conducted by Takaoka, Okano and Saijo (2011) with the permission of authors.

statistically significant (p -value < 0.001 , Wilcoxon rank-sum test)³⁵. Hence, MCM has strong effect making subjects more cooperative.

Takaoka, Okano and Saijo (2011) conducted a series of experiment with a session in which 22 subjects played *PDMC* for the first ten periods and then *PD* for the last ten periods, and a session in which another 22 subjects played *PD* for the first ten period and then *PDMC* for the last ten periods. Using the data in the first ten periods of these sessions in order to avoid possible order effects of *PD* and *PDMC*, they found that *PDMC* achieves 95.9% of the cooperation rate and *PD* achieves 7.7% of the cooperation rate. This difference is statistically significant (p -value < 0.001 , Wilcoxon rank-sum test)³⁶.

Observation 1.

(i) In the *PDMC* session, all twenty subjects chose the cooperative strategy in the dilemma stage, and then approved the other's choice in the mate choice mechanism.

(ii) In the *PD* game only session, the number of cooperation choice is 30 out of 380 (7.9%) and no (C,C) was observed among 190 pairs of choices.

(iii) The cooperation rate in the *PDMC* session is significantly different from that in the *PD* game only session.

9.2. The Robustness of the Mate Choice Mechanism

Let us describe the results of the *PDUV* session. Overall, the cooperation rate is 98.2%. For the early periods, some subjects chose the defection. At period 3 full cooperation was achieved for the first time. After period 11 all subjects chose the cooperative strategy. The difference of cooperation rates between *PDUV* and *PD* sessions is statistically significant (p -value < 0.001 , Wilcoxon rank-sum test). Hence, the unanimous voting also has strong effect making subjects more cooperative. The difference of cooperation rates between *PDUV* and *PDMC* sessions is not statistically significant at the 5 percent level (p -value = 0.076, Wilcoxon rank-sum test). This indicates that the difference of the wordings for the mechanism does not have a significant effect on the subjects' behavior.

Next, we describe the results of the sessions without repetition. All twenty subjects chose (C,y) in the *PDMC** session, and two subjects chose C and eighteen chose D in the *PD**

³⁵ When performing Wilcoxon rank-sum test, we first calculate average cooperation rate of each subject across periods, and then calculate the test statistic using the averages in order to eliminate correlation across periods.

³⁶ Using the data in the last ten periods, Takaoka, Okano and Saijo (2011) found that *PDMC* and *PD* achieve 95.9% and 9.1% of the cooperation rates respectively. The difference is statistically significant (p -value < 0.001 , Wilcoxon rank-sum test).

session (the cooperation rate is 10 percent), which is similar to the data of the *PDMC* and *PD* sessions. The difference of cooperation rates between *PDMC** and *PD** sessions is statistically significant (p -value < 0.001 , chi-square test)³⁷. Hence, the number of repetition does not have the effect on the performance of the mechanism. Summarizing the results, we have,

Observation 2.

- (i) *In the PD game with unanimous voting session, the cooperation rate is 98.2% and almost all defections occurred in the first few periods.*
- (ii) *The cooperation rate in the PD game with unanimous voting is significantly different from those in the PD game only session and not significantly different from those in the PDMC session.*
- (iii) *In the PD game with the mate choice mechanism session without repetition, all twenty subjects chose the cooperative strategy in the dilemma stage, and then approved the other's choice.*
- (iv) *In the PD game only session without repetition, two out of twenty subjects (10%) chose the cooperative strategy.*
- (v) *The cooperation rate in the PD game with the mate choice mechanism without repetition is significantly different from those in the PD game only session without repetition.*

9.3. The Comparison of the Mate Choice Mechanism with the Compensation Mechanism

Let us describe the results of the *CMPD* session. Overall, 76.6% of choices are cooperative, 68% for the first five periods increasing to 85% for the last five periods. The cooperation rate significantly increases as period passes, though it does not reach the full cooperation at period 19.³⁸ The cooperation rate is significantly different from that in the *PD* session (p -value < 0.001 , Wilcoxon rank-sum test), indicating that the compensation mechanism has the effect of making subjects more cooperative. The cooperation rate is also significantly different from those in the *PDMC* and *PDUV* sessions (p -value < 0.001 for both tests, Wilcoxon rank-sum test), indicating that the mate choice mechanism outperforms the compensation mechanism.

In the *SPE* and *BEWDS*, players should offer 3 or 4 (300 or 400 in the experiment). We find that the actual behavior is consistent with this prediction on average. Overall, the average side payment is 356.58. In each of 19 periods, subjects offer their side payment between 300 and 400 on

³⁷ In the chi-square test, the true cooperation rate in each session is replaced by its maximum likelihood estimate. Test statistic is distributed asymptotically as a chi-square with 1 degree of freedom under the null hypothesis.

³⁸ In order to examine whether the cooperation rate increases as period passes, we ran a simple random effect probit model. The dependent variable takes the value of 1 if the subject chooses C and 0 otherwise. The independent variables are the period number and the constant. The result indicates that the coefficient of the period is significantly greater than zero at the 1% significant level.

average. The minimum average side payment is 325 in period 6, and the maximum average side payment is 395 in periods 8 and 11. The *SPE* and *BEWDS* outcomes (3,3,C,C), (3,4,C,C) (or (4,3,C,C)), and (4,4,C,C) occurred 14, 27 and 25 times out of 190 outcomes respectively. The *BEWDS* outcomes (3,3,C,D) (or (3,3,D,C)), (3,3,D,D) and (3,4,C,D) (or (4,3,D,C)) occurred 0, 0 and 15 times out of 190 outcomes respectively.

Observation 3.

- (i) *In the CMPD session, 76.6% of choices are cooperative: 68% for the first five periods and 85% for the last five periods, and the cooperation rate significantly increases as period passes.*
- (ii) *The cooperation rate in the CMPD session is significantly different from those in the PD game only session, the PDMC session and the PDUV session.*
- (iii) *The average side payment is 356.58 with 325 as the minimum and 395 as the maximum where the subgame perfect side payment is between 300 and 400.*

9.4. Response to the outcome in the PD stage

Let us now consider how subjects respond to the outcome in the *PD* stage in the *PDMC* and *PDUV* sessions. First, all 190 pairs in the *PDMC* session chose (C,C) in the *PD* stage and all accepted the other choice in the approval stage. Second, 183 pairs in the *PDUV* session chose (C,C) in the *PD* stage and all except one accepted the other choice in the approval stage. In the questionnaire after the experiment the subject who chose the rejection wrote that he got bored with the succession of (C,C,y,y) outcomes. This subject chose defection in the *PD* stage in periods 4, 5, 10 and 11. Other two subjects chose defection in the *PD* stage. One chose defection only in period 1. Looking at the questionnaire, it seems that this subject did not consider the game deeply enough. Another one chose defection in periods 1 and 2. This subject thought that there might be some subjects who misunderstood the game rule at the beginning of the experiment, and might accept her defection. Therefore, there were seven choices of *D* in total, and all fell into the (C,D) case. Among these seven, subjects chose *C* always rejected the *D* choice of the other. On the other hand, subjects chose *D* always accepted the *C* choice of the other.

Observation 3.

- (i) *In the PDUV session, 183 pairs chose (C,C) and then all but one subject accepted the other choice in the decision stage.*
- (ii) *In the PDUV session, 7 pairs chose (C,D), subjects who chose C always rejected the D choice of the other and subjects chose D always accepted the C choice of the other in the decision stage.*

10. BEWDS and the Role of Mate Choice Flat

Our experimental results showed that almost all subjects chose the (C,C,y,y) path: 190 out of 190 pairs in the *PDMC* session and 182 out of 190 in *PDUV* session (7 pairs chose (D,C,y,n) and a pair chose (C,C,y,n) in the *PDUV* session). As shown in Properties 1 and 6, *SPE* (and hence Nash equilibria) have extra equilibrium paths other than (C,C,y,y) . On the other hand, the strategy pairs under *NSS* are $((C,y,n,\cdot), (C,y,n,\cdot))$ if both are payoff maximizers as Property 3 shows. Consider an off equilibrium path where subject 1 chooses D and subject 2 chooses C . Subject 1 has freedom to choose either y or n at this path since the fourth component of subject 1's strategy is either y or n under *NSS*, while subject 2 must choose n due to the third component. Therefore, both (n,n) and (y,n) are the candidates in the second stage or subgame DC if (D,C) is chosen in the first stage under *NSS*. In the *PDUV* session, we observed that seven pairs chose (D,C) in the first stage and then all chose (y,n) in the second stage. That is, no (n,n) was observed. On the other hand, consider three cases: both are payoff maximizers or reciprocators, or either one of the players is a reciprocator. Then, since the equilibrium strategy under *BEWDS* is either (C,y,n,y,\cdot) or (C,y,n,y,n) , (y,n) must be chosen at subgame DC (subject 1 must choose y in the fourth component of either strategy, and subject 2 must choose n in the third component). Therefore, although the number of pairs is small, *BEWDS* most exactly predict the subjects' behavior in *PDUV* sessions.

It is difficult to detect whether the subjects adopted the norm of reciprocity since no (D,D) is observed. The difference between a payoff maximizer and a reciprocator under *BEWDS* is the action played at subgame DD : the payoff maximizer chooses either y or n while the reciprocator chooses n . Furthermore, most subjects did not mention what they would have chosen at subgame (D,D) in the questionnaire. For this reason, we will restrict ourselves to *BEWDS* in the following analysis.

In the first half of this section, we will consider why *BEWDS* has more predictive power on subjects' behavior than *NE*, *SPE* or *NSS* when the *MCM* is employed. A possible reason is that there are simple heuristics whose strategies are equivalent to *BEWDS*. Furthermore, questionnaire analysis finds that most subjects were likely to adopt them. In the second half of this section, we will argue that, with subjects behaving in line with *BEWDS*, the *MCM* has the uniqueness property under some axioms.

The mate choice flat derives from the *MCM*, and it has several nice features. Although the following property is trivial, the mate choice flat alleviates cognitive burden of subjects from two dimensional to one dimensional comparison.

Property 8. Under the mate choice flat, strategy α with the payoff vector (u,v) weakly dominates strategy β with (x,y) if and only if $u > x$.

Since subjects can instantaneously understand that each subgame in the second stage has the mate choice flat due to the MCM, and can identify (10,10) as the outcome of subgame DD in Figure 7, 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. Subject 2's outcomes to be compared are the two upper right triangles of the upper row. Hence, the number of triangles that each subject must see is four in each of subgames CC, CD and DC. Hence, the remaining triangles are unnecessary for their decision making including the lower right cell.

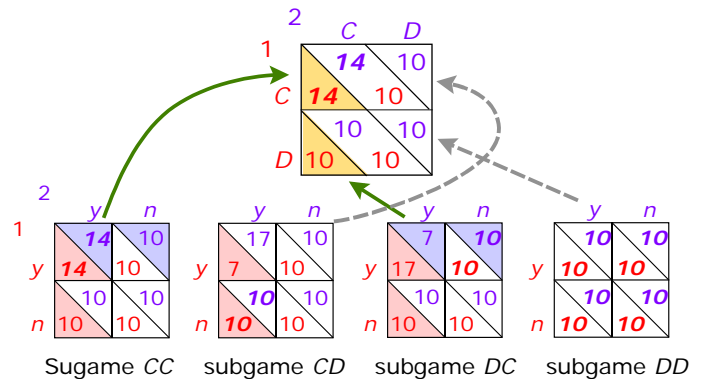


Figure 7. The Triangles that Subjects must Consider and Backwardability.

Let us consider the minimum informational or intellectual requirement to achieve (C,C,y,y) 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 7. 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, and we found many subjects actually employ the methods in the following.³⁹ If a subject is an *R*, since the decisions in subgames have already determined, the cognitive burden is just to compare 14 with 10 in the reduced normal form game.

Subjects in *PDMC* and *PDUV* sessions answered written questionnaires during and after the experiment. For example, they were asked the following questions:

a) During the experiment,

- 1) Describe your way of thinking when you decide the choice.
- 2) Describe your way of thinking when you decide to accept or reject (agree or disagree).

b) After the experiment,

- 1) Were there something that you found about how to determine your and your partner's earning points? In particular, what did you refer when you decided your choice and decision (voting)?
- 2) Were there something that you found about the action of paired person in the choice stage? Knowing that, what did you aim to respond in the later choice stage?
- 3) What did you refer when you either accept or reject (agree or disagree) the action of paired person in the decision (voting) stage?
- 4) Were there something that you found about the decision of paired person in the decision (voting) stage? Knowing that, what did you aim to respond in the following period?
- 5) Looking back, what do you think is the best play in this experiment? And why?

³⁹ Player 1 must compare two numbers six times in order to decide (*C,y*) in Figure 7: two comparisons (i.e., my own 14 and 10, and the other's 14 and 10) in subgame *CC*, one comparison (i.e., my own 10 and 7, and the other's choice does not matter since my own outcome is 10 regardless the choice of the other) in subgame *CD*, two comparisons in subgame *DC*, and one comparison between 14 and 10 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*.

From these questionnaires, we will detect whether subjects adopt backwardability, whether subjects adopt the idea of weak dominance or its equivalence (Property 8) and whether subjects adopt simple heuristic that “*the other subject thinks the same way as I think*”.

	Backwardability	Weak Dominance or Its Equivalence	The Same as I Think
Complete Description	20	19	16
Partial Description	0	1	3
No Description	0	0	1

Table 2: The *PDMC* session

	Backwardability	Weak Dominance or Its Equivalence	The Same as I Think
Complete Description	20	17	14
Partial Description	0	3	5
No Description	0	0	1

Table 3: The *PDUV* session

For counting the questionnaire, we recruited an economics graduate student who did not know the contents of this project at all in order to maintain objectivity. He received a written instruction for counting in which the summary of the experiment was written, which is the same as that the subjects were received in the experimental session, and he was asked to count the number of subjects who seemed to have in mind the concept of backwardability, weak dominance or its equivalence and a heuristic that “*the other subject thinks the same way as I think*”. He was not provided the decision criteria for counting and was asked to set up it by himself.

Tables 2 and 3 show the results of counting. In both *PDMC* and *PDUV* sessions, most subjects described that they had the idea of backwardability, weak dominance or its equivalence and the simple heuristic that “*the other subject thinks the same way as I think*.” In every component, more than or equal to 70 percent of subjects describe completely the corresponding idea in both sessions. Furthermore, 16 out of 20 subjects describe completely all three components in *PDMC* session, and 13 in *PDUV* session. Note that our questionnaire was free description. We did not restrict subjects such that they would pay attention to these components. This analysis indicates that subjects seem to behave in line with *BEWDS* under the *MCM*.

MCM that we employed has the uniqueness property under *BEWDS*. That is, any approval mechanisms satisfying Axioms 1, 2 and 3 in the following must be *MCM*. We say that an approval mechanism satisfies *forthrightness* if both choose y in the second stage after the choice

of a strategy pair in a *PD* game, then the outcome of the approval mechanism is the outcome of the *PD* game with the strategy pair.⁴⁰ For example, suppose that subjects 1 and 2 choose (C,D) and both choose y in the approval mechanism. Then forthrightness requires that the outcome must be (C,D) . In order to limit the class of approval mechanisms, we introduce the following axioms.

Axiom 1 (*Onto*): An approval mechanism satisfies the *onto* condition if every outcome of a *PD* game is an outcome of *PDMC* and every outcome of the *PD* game with the approval mechanism must be an outcome of the prisoner's dilemma game.

The onto condition requires that the set of outcomes of subgames *CC*, *CD*, *DC* and *DD* must be $\{(14,14),(7,17),(17,7),(10,10)\}$. Consider a (non-approval) mechanism that gives 5 if a subject chooses "C" in Figure 1 as reward for cooperation. If this reward mechanism is employed, then the outcomes become $(14+5,14+5),(7+5,17),(17,7+5)$, and $(10,10)$. That is, $(14,14),(7,17),(17,7)$ would never be realized with this mechanism. Therefore, this reward mechanism with the *PD* game does not maintain the outcomes of the *PD* game. In this sense, many reward or punishment mechanisms are not "onto". The onto condition also excludes mechanisms that are not budget balanced. That is, the reward mechanism above needs outside money to maintain the mechanism, and the onto condition does not allow this budget deficit. Furthermore, punishment occasionally reduces total payoff, and hence it is not efficient.

Axiom 2 (*Mate choice flat at the approval stage*): An approval mechanism satisfies *mate choice flat at the approval stage* if either subject chooses n , the outcome of these strategy pairs must be the same for each subgame.

Axiom 2 allows that the flat outcome in subgame *CC* can be different from the one of subgame *CD*.

Axiom 3 (*Mate choice flat at the reduced normal form stage*): A *PD* game with an approval mechanism satisfies *mate choice flat at the reduced normal form stage* if either subject says "D" in the normal form game derived from the two stage game, the outcome of these strategy pairs must be the same.

⁴⁰ This definition is slightly different from forthrightness introduced by Tatsuyoshi Saijo, Yoshikatsu Tatamitani and Takehiko Yamato (1996).

October 14, 2011
Not for circulation!

This axiom requires that the reduced normal form game of the two stage game must also have a mate choice flat if either subject chooses D , but does not require that the same outcome is related to the outcomes in the second stage.

We say that an approval mechanism with a PD game is *natural* if the outcome is $(10,10)$ when either one of two subjects chooses n , and it is *voluntary* if any subject who chooses D should not be forced to change from D to C . Clearly, if it is natural, it should be voluntary since a defector is not forced to contribute \$10. The forthright and natural mechanism is exactly MCM by its construction. It is straightforward to see that the forthright and natural mechanism satisfies Axioms 1, 2 and 3. On the other hand, the following property guarantees that a forthright mechanism satisfying Axioms 1, 2 and 3 must be natural, and hence it is voluntary.

Property 9. *Suppose that the unique equilibrium path of a PD game with an approval mechanism is (C,C,y,y) under $BEWDS$ and suppose Axioms 1, 2 and 3. Then the approval mechanism satisfying forthrightness is natural.*

Proof. See Appendix.

Since an approval mechanism that is natural and forthright must be MCM , Property 9 shows the uniqueness of the mechanism under Axioms 1, 2 and 3.

11. Concluding Remarks

We found that the MCM promotes cooperation significantly in the PD game. This experimental evidence is most compatible with the behavioral principle based upon $BEWDS$ including reciprocal behavior. The elimination of weakly dominated strategies can be done by comparison of two numbers, but not two vectors due to the mate choice flat. It seems that the flat reduced subjects' cognitive burden, and made them easily consider backwardly. We noticed that the MCM is unique with several axioms. We also found that the cooperation rate in $PDMC$ is significantly higher than that in $CMPD$.

Of course, the MCM does not always solve all prisoner's dilemma. First, the participants must agree upon using the mechanism as mechanism designers of all fields in economics implicitly presume. Second, the mechanism might need monitoring devices and/or enforcing power. Otherwise, a participant might not conduct the deed described in " C " even after two participants choose " C " and " y ". Third, we cannot apply the mechanism if the contents of " C " have not been settled down before applying it. Many researchers use global warming as an example of PD . Although countries and parties have been negotiating the substance of coping

with it for over twenty years under the United Nations Framework Convention on Climate Change (*UNFCCC*), they have not reached what exactly "C" should be.

The *PD* game has two participants and two strategies. We will consider the directions of our further research agenda based upon these numbers. First, fix the number of participants two, and then consider the number of strategies is at least three. This is nothing but a voluntary contribution mechanism for the provision of a public good with two participants. Masuda, Okano and Saijo (2011) show that the *MCM* with *BEWDS* cannot implement the Pareto outcome when both have the same linear utility function. Then they designed the minimum mate choice mechanism that is based upon the spirit of the *MCM*, and found that it implements the Pareto outcome theoretically and experimentally. The contribution rates of several sessions exceeded 95%.

Second, fix the number of strategies two, and then consider the number of participants is at least three. This is nothing but a social dilemma situation. As Banks, Plott and Porter (1988) found, Okano, Masuda and Saijo (2011) also show that the *MCM* with *BEWDS* cannot implement the Pareto outcome. Then they design new mechanisms utilizing the idea of the *MCM* that implement the Pareto outcome. However, the cooperation rates in the experiment are about 70%. This is due to the fact that the mechanisms implementing the Pareto outcome necessarily contain *PD* games with two participants. For example, consider the case where one player chooses *D*, and the other two players choose *C*. In the second stage, two players with choice *C* face a *PD* game. Two players should not cooperate theoretically in order to attain full cooperation (i.e., the three choose *C*), but they occasionally choose *C* experimentally. In other words, cooperation of two players is a major stumbling block against full cooperation that is a fundamental difficulty in designing workable mechanisms in social dilemma. Although many researchers do not see any differences between two and more than two participants, they find a deep fissure between them.

Third, consider that both are at least three. This environment is a wide open area. Of course, there are quite a number of papers with this environment (see, for example, related papers in Plott and Vernon L. Smith (2008)), but the crack between theory and experiment has not been filled up.

Mechanism designers have not been considering *comfortability* of mechanisms. Although it is still early stage of research, Hideo Shinagawa, Masao Nagatsuka, Okano and Saijo (2011) find that subjects facing the *MCM* did not show any significant activation of anterior prefrontal cortex (*PFC*) in the processing of decision making using a near-infrared spectroscopy (*NIRS*)-based system. This finding suggests the possibility that subjects with the *MCM* made decision at ease. On the other hand, subjects playing the *PD* game showed significant activation of right *PFC* and

left orbitofrontal cortex that are related to unpleasant emotion.⁴¹

Takaoka, Okano and Saijo (2011) compare the *MCM* with costly punishment measuring salivary alpha-amylase (*sAA*) of subjects. *SAA* has been proposed as a sensitive biomarker for stress-related changes in the body that reflect the activity of the sympathetic nervous system. They find that subjects who experienced the *MCM* reduced the level of *sAA* and subjects who experienced costly punishment increased the level of *sAA*. This indicates that the *MCM* is a mechanism that is relatively stress free.

Appendix

	Engineering	Engineering Science	Economics	Letters	Law	Foreign Studies	Human Sciences	Science	Frontier Biosciences	Medicine	Pharmaceutical Sciences	Dentistry	# of females	Average age	Average Payoff (\$)	Max Payoff (\$)	Min Payoff (\$)
<i>PD</i>	4	7	2	2	1	1		1	1		1		3	22.6	45.4	48.9	30.8
<i>PDMC</i>	5	4	3		4	1	1			1	1		4	22.1	61.6	61.6	61.6
<i>PDUV</i>	4	4	2		3	2	1	1		2		1	9	21.4	59.3	60.0	55.5
<i>CMPD</i>	2	4	1	2		2	2	5		1	1		6	21.4	56.1	58.7	53.0
<i>PD*</i>	12	2	1		1		2	1		1			3	23.0	48.1	78.6	32.4
<i>PDMC*</i>	11	2	1	1	2	1	2						2	21.8	65.1	65.1	65.1

1) Numbers of divisions show the numbers of participants.

2) No repetition in "*" sessions.

3) \$1=86.3 yen for *PD* and *PDMC*, \$1=88.6 yen for *PDUV*, *CMPD*, \$1=82.2 for *PD** and \$1=81.8 for *PDMC**.

Table A1. Subjects' Information

Proof of Property 3. Let $P(s_1, s_2)$ be the path with (s_1, s_2) . The following four cases cover all strategies.

Case 1. $t = (D, \cdot, \cdot, \cdot)$ is not an *NSS*.

Let $t' = (C, y, n, n, \cdot)$. Since $P(t, t) = (D, D, \cdot, \cdot)$, $P(t', t) = (C, D, n, \cdot)$, $P(t, t') = (D, C, \cdot, n)$ and $P(t', t') = (C, C, y, y)$, we have $v(t, t) = v(t', t) = v(t, t') = 10$ and $v(t', t') = 14$. Therefore, t is not an *NSS*.

Case 2. $t = (C, n, \cdot, \cdot)$ is not an *NSS*.

Let $t' = (C, y, \cdot, \cdot)$. Since $P(t, t) = (C, C, n, n)$, $P(t', t) = (C, C, y, n)$, $P(t, t') = (C, C, n, y)$ and $P(t', t') = (C, C, y, y)$, we have $v(t, t) = v(t', t) = v(t, t') = 10$ and $v(t', t') = 14$. Therefore, t is not an *NSS*.

Case 3. $t = (C, y, y, \cdot)$ is not an *NSS*.

⁴¹ See Yoko Hoshi, Jinghua Huang, Shunji Kohri, Yoshinobu Iguchi, Masayuki Naya, Takahiro Okamoto, and Shuji Ono (2011).

Let $t' = (D, \cdot, \cdot, y, \cdot)$. Since $P(t, t) = (C, C, y, y)$ and $P(t', t) = (D, C, y, y)$, we have $v(t, t) = 14 < v(t', t) = 17$. Therefore, t is not an NSS.

Case 4. $t = (C, y, n, \cdot, \cdot)$ is an NSS.

Since $P(t, t) = (C, C, y, y)$, $v(t, t) = 14$.

(i) If $t' = (D, \cdot, \cdot, \cdot, \cdot)$, $v(t', t) = 10$ because $P(t', t) = (D, C, \cdot, n)$. Therefore, $v(t, t) = 14 > v(t', t) = 10$.

(ii) If $t' = (C, n, \cdot, \cdot, \cdot)$, $v(t', t) = 10$ because $P(t', t) = (C, C, n, y)$. Therefore, $v(t, t) = 14 > v(t', t) = 10$.

(iii) If $t' = (C, y, \cdot, \cdot, \cdot)$ with $t' \neq t$, $P(t', t) = P(t, t') = P(t', t') = (C, C, y, y)$. Therefore, $v(t', t) = v(t, t') = v(t', t') = 14$. Hence, $v(t, t) = v(t', t)$ and $v(t, t') = v(t', t')$.

These cases show that t is an NSS. ■

Proof of Property 9. We will show a slightly general proof allowing asymmetry of the PD game matrix. Consider the following PD game in Figure A1. Each cell represents (player 1's payoff, player 2's payoff). We assume that $c > a > d > b$ and $x > w > z > y$.

	C	D
C	(a,w)	(b,x)
D	(c,y)	(d,z)

Figure A1. Prisoner's dilemma game.

First, consider the second stage, i.e., the approval mechanism stage. In Figures A2, A3, A4 and A5, the upper choice is for y and the lower is for n for player 1 and the left is for y and the right is for n for player 2. Consider subgame CC. Then the upper left cell must be (a, w) by forthrightness in Figure A2, and there are four possibilities of the flat by Axioms 1 and 2.

(a,w)	(a,w)	(a,w)	(b,x)	(a,w)	(c,y)	(a,w)	(d,z)
(a,w)	(a,w)	(b,x)	(b,x)	(c,y)	(c,y)	(d,z)	(d,z)
(1)	(2)	(3)	(4)				

Figure A2. Four Possible Cases at subgame CC.

Shaded areas show the remaining outcomes using elimination of weakly dominated strategies. Since (C, C, y, y) is the unique path, (4) must be the case among the four possibilities.

Applying the same procedure for subgames CD, DC and DD, we have the following figures.

(b,x)	(b,x)	(b,x)	(a,w)	(b,x)	(c,y)	(b,x)	(d,z)
(b,x)	(b,x)	(a,w)	(a,w)	(c,y)	(c,y)	(d,z)	(d,z)
(1)	(2)	(3)	(4)				

Figure A3. Four Possible Cases at subgame CD.

(c,y)	(c,y)
(c,y)	(c,y)

(1)

(c,y)	(a,w)
(a,w)	(a,w)

(2)

(c,y)	(b,x)
(b,x)	(b,x)

(3)

(c,y)	(d,z)
(d,z)	(d,z)

(4)

Figure A4. Four Possible Cases at subgame DC.

(d,z)	(d,z)
(d,z)	(d,z)

(1)

(d,z)	(a,w)
(a,w)	(a,w)

(2)

(d,z)	(b,x)
(b,x)	(b,x)

(3)

(d,z)	(c,y)
(c,y)	(c,y)

(4)

Figure A5. Four Possible Cases at subgame DD.

Second, consider the reduced normal form game stage. The outcome of the upper left cell with (C,C) must be (a,w) in each game of Figure A6, and the cells of the rest must have the same outcome in each game due to Axioms 1 and 3. The shaded areas in Figures A6 show the outcomes of elimination of weakly dominated strategies.

(a,w)	(a,w)
(a,w)	(a,w)

(1)

(a,w)	(b,x)
(b,x)	(b,x)

(2)

(a,w)	(c,y)
(c,y)	(c,y)

(3)

(a,w)	(d,z)
(d,z)	(d,z)

(4)

Figure A6. Four Possible Mate Choice Flats in the Reduced Normal Form Game.

Finally, since (C,C,y,y) is the unique path, (4) is the only possible case in Figure A6. That is, the outcome of mate choice flat is (d,z) in the reduced normal form game. Let us go back to Figures A3, A4 and A5 where the outcome (d,z) must be chosen (i.e., (4) in Figure A3, (4) in Figure A4 and (1) in Figure A5). Then (d,z) is the outcome of mate choice flat for each case. Since the outcome (d,z) is also the mate choice flat in Figure A2, all four cases have the common mate choice flat (d,z) in the approval mechanism. Hence, this approval mechanism must be natural. ■

References

- Andreoni, James, and John H. Miller. 1993. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence." *Economic Journal*, 103(418): 570-85.
- Andreoni, James, and Hal Varian. 1999. "Preplay Contracting in the Prisoners' Dilemma." *Proceedings of the National Academy of Sciences*, 96(19): 10933-8.
- Aumann, Robert J. 2006. "War and Peace." *Proceedings of the National Academy of Sciences*, 103(46): 17075-78.
- Banks, Jeffrey S., Charles R. Plott, and David P. Porter. 1988. "An Experimental Analysis of Unanimity in Public Goods Provision Mechanisms." *Review of Economic Studies*, 55(2): 301-22.
- Bereby-Meyer, Yoella, and Alvin E. Roth. 2006. "The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation." *American Economic Review*, 96(4): 1029-42.
- Cason, Timothy N., Tatsuyoshi Saijo, and Tomas Sjöström, and Takehiko Yamato. 2006. "Secure Implementation Experiments: Do Strategy-proof Mechanisms Really Work?" *Games and Economic*

October 14, 2011
Not for circulation!

Behavior, 57(2): 206-235.

Charness, Gary, Guillaume R. Fréchet, and Cheng-Zhong 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.

Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. 1996. "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games." *Games and Economic Behavior*, 12(2): 187-218.

Croson, Rachel T.A. 2007. "Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games." *Economic Inquiry*, 45(2): 199-216.

Doebeli, Michael, and Christoph Hauert. 2005. "Models of Cooperation Based on the Prisoner's Dilemma and the Snowdrift Game." *Ecology Letters*, 8(7): 748-766.

Duffy, John, and Jack Ochs. 2009. "Cooperative Behavior and the Frequency of Social Interaction." *Games and Economic Behavior*, 66(2): 785-812.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review*, 90(4): 980-94.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics*, 10(2): 171-8.

Flood, Merrill M. 1958. "Some Experimental Games." *Management Science*, 5(1): 5-26.

Gächter, Simon, and Christian Thöni. 2005. "Social Learning and Voluntary Cooperation among Like-Minded People." *Journal of the European Economic Association*, 3(2-3): 303-14.

Guala, Francesco. 2010. "Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate." University of Milan Department of Economics, Business and Statistics Working Paper No. 2010-23. Forthcoming in *Brain and Behavioral Sciences*.

Halliday, T. R. 1983. "The Study of Mate Choice." In *Mate Choice*, ed. Patrick Bateson, 3-32. New York: Cambridge University Press.

Hamilton, W. D.. 1964. "The Genetical Evolution of Social Behaviour." *Journal of Theoretical Biology*, 7(1): 1-16.

Hauert, Christoph, Arne Traulsen, Hannelore Brandt, Martin A. Nowak, and Karl Sigmund. 2007. "Via Freedom to Coercion: The Emergence of Costly Punishment." *Science*, 316(5833): 1905-7.

Hoshi, Yoko, Jinghua Huang, Shunji Kohri, Yoshinobu Iguchi, Masayuki Naya, Takahiro Okamoto, and Shuji Ono. 2011. "Recognition of Human Emotions from Cerebral Blood Flow Changes in the Frontal Region: A Study with Event-Related Near-Infrared Spectroscopy." *Journal of Neuroimaging*, 21(2): 94-101.

Hume, David. 1739 (1874 Edition). *A Treatise of Human Nature Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects and Dialogues Concerning Natural Religion*. London: Longmans, Green, and Co.

October 14, 2011
Not for circulation!

- Jackson, Matthew O. 2001. "A Crash Course in Implementation Theory." *Social Choice and Welfare*, 18(4): 655-708.
- Kalai, Ehud. 1981. "Preplay Negotiations and the Prisoner's Dilemma." *Mathematical Social Sciences*, 1(4): 375-9.
- Kandori, Michihiro. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies*, 59(1): 63-80.
- Kreps, David M., Paul Milgrom, John Roberts, and Robert Wilson. 1982. "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory*, 27(2): 245-52.
- López-Pérez, Raúl, and Marc Vorsatz. 2010. "On Approval and Disapproval: Theory and Experiments." *Journal of Economic Psychology*, 31(4): 527-41.
- Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. Oxford: Oxford University Press.
- Maskin, Eric. 1999. "Nash Equilibrium and Welfare Optimality." *Review of Economic Studies*, 66(1): 23-38.
- Masuda, Takehito, Yoshitaka Okano and Tatsuyoshi Saijo. (2011). "The Minimum Approval Mechanism implements Pareto Efficient Outcome Theoretically and Experimentally." In Preparation.
- Moore, John, and Rafael Repullo. 1988. "Subgame Perfect Implementation." *Econometrica*, 56(5): 1191-1220.
- Nakamaru, Mayuko, Tatsuyoshi Saijo, and Takehiko Yamato. 2011. "Replicator Dynamic Model of Prisoner's Dilemma Game with the Mate Choice Mechanism." In Preparation.
- Nowak, Martin A. 2006. "Five Rules for the Evolution of Cooperation." *Science*, 314(5805):1560-1563.
- Okano, Yoshitaka, Tatsuyoshi Saijo, and Junyi Shen. 2011. "Backward Elimination of Weakly Dominated Strategies is Compatible with Experimental Data Compared with Subgame Perfection: Approval and Compensation Mechanisms." In Preparation.
- Ostrom, Elinor. 1990. *Governing the Commons*. Cambridge: Cambridge University Press.
- Plott, Charles R. and Vernon L. Smith. 2008. *Handbook of Experimental Economics Results*. Amsterdam: North-Holland.
- Poundstone, William. 1992. *Prisoner's Dilemma*. New York: Anchor.
- Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5): 1281-1302.
- Roth, Alvin E. 1995. "Introduction to Experimental Economics." In *The Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth, 3-109. Princeton: Princeton University Press.

October 14, 2011
Not for circulation!

- Roth, Alvin E. and J. Keith Murnighan. 1978. "Equilibrium Behavior and Repeated Play of the Prisoner's Dilemma." *Journal of Mathematical Psychology*, 17(2): 189-98.
- Russett, Bruce, Harvey Starr and David Kinsella. 2009. *World Politics: The Menu for Choice*, Wadsworth Publishing. 9th edition.
- Wadsworth Publishing; 9 edition (January 16, 2009)
- Saijo, Tatsuyoshi . 1988. "Strategy Space Reduction in Maskin's Theorem: Sufficient Conditions for Nash Implementation." *Econometrica*, 56(3): 693-700.
- Saijo, Tatsuyoshi, and Yoshitaka Okano. 2009. "Six Approval Rules Whose Outcomes are Exactly the Same as the Mate Choice Rule." mimeo.
- Saijo, Tatsuyoshi, Yoshikatsu Tatamitani, and Takehiko Yamato. 1996. "Toward Natural Implementation." *International Economic Review*, 37(4): 949-80.
- Saijo, Tatsuyoshi, Tomas Sjöström, and Takehiko Yamato. 2007. "Secure Implementation." *Theoretical Economics*, 2(3): 203-229.
- Selten, R. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory*, 4(1): 25-55.
- Shinagawa, Hideo, Masao Nagatsuka, Yoshitaka Okano and Tatsuyoshi Saijo. 2011. "Cerebral Blood Flow Changes in the Frontal Region of Approval Mechanism and Prisoner's Dilemma Game: A Study with Event-Related Near-Infrared Spectroscopy." In Preparation.
- Shubik, Martin. July 2011. "The Present and Future of Game Theory," Cowles Foundation Discussion Paper 1808, Yale University.
- Smith, Adam. 1759. *The Theory of Moral Sentiments*. Glasgow, Scotland.
- Sugden, Robert. 1984. "The Supply of Public Goods Through Voluntary Contributions." *Economic Journal*, 94(376): 772-787.
- Takaoka, Masanori, Yoshitaka Okano, and Tatsuyoshi Saijo. 2011. "Institutional Stress between Approval and Costly Punishment Mechanisms: A Salivary Alpha-Amylase Approach." In Preparation.
- Varian, Hal R. 1994. "A Solution to the Problem of Externalities When Agents Are Well-Informed." *American Economic Review*, 84(5): 1278-93.
- Yamagishi, Toshio. 1986. "The Provision of a Sanctioning System as a Public Good." *Journal of Personality and Social Psychology*, 51(1): 110-116.