Discussion Paper No. 874

THE MINIMUM APPROVAL MECHANISM IMPLEMENTS THE EFFICIENT PUBLIC GOOD ALLOCATION THEORETICALLY AND EXPERIMENTALLY

Takehito Masuda
Yoshitaka Okano
Tatsuyoshi Saijo

May 2013

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

The Minimum Approval Mechanism Implements the Efficient Public Good Allocation Theoretically and Experimentally

November 2012

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

[a] *Graduate School of Economics, Osaka University 1-7, Machikaneyama, Toyonaka, Osaka 560-0043, Japan*

[b] *Research Center for Social Design Engineering, Kochi University of Technology Tosayamada, Kami-city, Kochi 782-8502, Japan*

[c] *Center for Environmental Innovation Design for Sustainability, Osaka University, 2-1 Yamada-oka, Suita, Osaka 565-0871, Japan*

Abstract

We propose the minimum approval mechanism (MAM) for a standard linear public good environment with two players. Players simultaneously and privately choose their contributions to the public good in the first stage. In the second stage, they simultaneously decide whether to approve the other's choice. Both contribute what they choose in the first stage if both players approve; otherwise, both contribute the minimum of the two choices in the first stage. The MAM implements the Pareto-efficient allocation in backward elimination of weakly dominated strategies (BEWDS) and is unique under plausible conditions. Contributions in the MAM experiment overall averaged 94.9%. The data support BEWDS rather than subgame perfect Nash equilibria. Quantifying subjects' responses to the questionnaire showed that the majority of subjects in the MAM found a heuristic or an algorithm named *diagonalization* and supported the notions of *minimax regret* and *iterated best response,* all of which mimic BEWDS outcomes.

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

Keywords: Public good experiment; Approval mechanism; Assumed equilibrium concepts

---

* Corresponding author. Postal address: Institute of Social and Economic Research, Osaka University, 6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan. Tel.: +81-6-6879-8582; Fax: +81-6-6879-8583. E-mail address: tmasuda@iser.osaka-u.ac.jp (T. Masuda). Appendices are available upon request.

Abbreviations: MAM = minimum approval mechanism, BEWDS = backward elimination of weakly dominated strategies, MCM = mate choice mechanism.

1. Introduction

Ever since the findings of Samuelson (1954) highlighted the inefficiency of market mechanism-based allocations, dominant strategy, Nash, and subgame perfect Nash have been the major equilibrium concepts that players are assumed to follow when theorists design public good mechanisms that implement the efficient allocation. For example, Green and Laffont (1977) characterized the class of dominant strategy mechanisms, while Groves and Ledyard (1977) theoretically designed a budget-balanced mechanism that implements the efficient allocation in Nash equilibrium outcomes.[1] Experimental research that emerged at around the same time began to evaluate the performances of these and other mechanisms.

Much experimental research has since focused on evaluating the performances of single-stage mechanisms with efficient dominant strategy equilibria or Nash equilibria (NE). Contrary to the theoretical stance put forward above, subjects have actually been found to frequently deviate from truthful dominant strategies and become stuck at weakly dominated NE (Attiyeh et al., 2000). To solve this problem, Saijo et al. (2007) introduced secure implementation (or double implementation in dominant strategy equilibria and NE). Further, Cason et al. (2006) experimentally showed that the frequency of dominant strategy play was significantly increased in a secure Groves–Clarke mechanism relative to a non-secure pivotal mechanism, although the frequency was only slightly over 80%.

By contrast, the Groves–Ledyard (1977) and Walker (1981) mechanisms have difficulty attaining efficient NE in one-shot play because of subjects' bounded rationality.[2] As a consequence, the papers by Chen and Gazzale (2004), Healy (2006), and Healy and Mathevet (2012) characterized the stable mechanisms necessary to achieve NE as rest points of various learning dynamics. Nonetheless, these studies still allow dozens of repetitions, which may not be applicable in practical situations. A notable exception is Falkinger et al. (2000), which used Falkinger's (1996) mechanism to observe that contributions approached NE in only a few periods.[3] However, this mechanism taxes or subsidizes proportionally to how much a player's contribution deviates from the mean contribution and hence some players are forced to contribute more than they choose.

Experimentalists have also assessed the empirical validity of multi-stage mechanisms. Based upon subgame perfect Nash equilibria (SPNE), Varian (1994) designed a two-stage compensation mechanism where players decide how much to subsidize other players before they make contributions; this mechanism was subsequently experimentally assessed by Andreoni and Varian (1999) using asymmetric prisoner's dilemma games. The

[1] For arguments in a more general environment, see Laffont and Maskin (1982).
[2] See Chen and Plott (1996) and Chen and Tang (1998).
[3] Contributions averaged 90.5% of subjects' endowment in the experiments with linear payoffs.

authors found that approximately two-thirds of subsidy offers were sufficiently high to incentivize the other player to cooperate, and by the end of the experiment 77.4% of subjects learned to choose cooperation for such offers.[4] However, Hamaguchi et al. (2003) observed systematic deviations from SPNE in emissions trading experiments using this two-stage compensation mechanism. These findings confirm that compensation mechanisms work well in simple environments but less so in complicated ones. Smith (1979, 1980) began to apply continuous auction processes but their theoretical foundations were not fully understood. Later, Banks et al. (1988) found that the trembling-hand perfect equilibrium allocation is efficient in the Smith auction process,[5] but their experimental results were mixed. They showed that although the Smith auction process significantly increased average contributions relative to the voluntary contribution mechanism (VCM), it repeatedly decreased contributions instead of converging to the equilibrium play.

In summary, previous work has shown that multi-stage mechanisms facilitate the efficient provision of public goods as well as single-stage mechanisms do, although the former require subjects to perform more complicated strategic considerations than the latter. Still, the theory is not fully in accordance with the experimental evidence, and this gap may come from the commonly used research method, namely that theorists assume one equilibrium concept a priori to construct public good mechanisms.

Motivated by this problem, Saijo et al. (2012) investigated several equilibrium concepts under the mate choice mechanism (MCM), which they developed to implement cooperative outcome in prisoner's dilemma games. In the MCM, each player chooses to approve or disapprove after observing the other player's choice (cooperation or defection). If both players approve, the outcome is what they chose in the first stage; otherwise, the outcome is defection for both. Noteworthily, this prisoner's dilemma experiment using the MCM achieved a 93.2% cooperation rate under perfect stranger matching. The authors also found that backward elimination of weakly dominated strategies (BEWDS) is more compatible with experimental data than the NE, SPNE or neutrally stable strategies are.[6]

This paper thus develops the minimum approval mechanism (MAM) that implements the Pareto-efficient allocation in BEWDS in linear public good environments with two players, in the spirit of the approach presented by Saijo et al. (2012). In the first stage of the MAM, players simultaneously and privately choose their contributions to the public good. In the second stage, players simultaneously decide whether to approve the other player's choice after observing it. If both players approve, each player's final

[4] Charness et al. (2007) showed that 29–60% of subsidy offers were consistent with the subgame perfect equilibrium in their prisoner's dilemma experiments with this compensation mechanism.

[5] This finding is based on the assumption that players bid once.

[6] As Saijo et al. (2012) noted, Kalai (1981) used BEWDS.

contribution is what he or she chooses in the first stage; otherwise, his or her final contribution is the minimum of the two choices in the first stage.

Moreover, we define the class of approval mechanisms in order to show that the MAM is a unique approval mechanism that satisfies the following three plausible properties, implementing the Pareto-efficient allocation in BEWDS. *Voluntariness* ensures that no player is forced to contribute more than he or she chooses in the first stage. In other words, we only use refunds to players rather than coercive power. This notion of voluntariness is a noteworthy feature of the MAM in contrast to tax/subsidy schemes such as the Groves–Ledyard and Falkinger mechanisms. *Monotonicity* states that players' contributions are weakly increasing in their first-stage choices. *Forthrightness* states that i) if both players make the same first-stage choice or both approve, then they contribute what they chose in the first stage and ii) given first-stage choices, disapproval by either player results in the same outcome.[7]

This paper makes the following contributions to the body of knowledge on this topic. In experiments under the perfect stranger matching protocol, we observe a rapid convergence to the efficient allocation in the MAM. Starting from 76.9%, subjects successfully sustained cooperation, reaching an overall average contribution rate of 94.9% of endowments. A similar result is obtained for the simplified MAM (SMAM) treatment, in which only the player that has the higher first-stage choice can proceed to the second stage. These results contrast with those of experimental studies that need repetition to achieve equilibrium (see Andreoni and Varian, 1999; Chen and Plott, 1996; Chen and Tang, 1998; Healy, 2006). We also conduct the MCM so that unilateral disapproval results in no contribution and find that this mechanism increases contributions to 28.4% from 10.2% for the baseline VCM.

We compare the predictive performances of BEWDS and SPNE using two criteria that deal with multiple equilibria and inclusion between the sets of equilibrium paths. This approach is different to that presented in the behavioral game theory literature (see Ert et al., 2011) where the distance between data and predictions is well defined. The first criterion relates to the low number of equilibrium paths, while the second is the proportion of equilibrium-consistent path data. Our evaluation method shows that BEWDS has a superior predictive performance compared with SPNE in the MAM and SMAM. However, we cannot conclude whether BEWDS or SPNE is more predictive in the MCM. By applying the same argument to the second-stage decisions (with a slight modification of subgame consistency based on the findings of Binmore et al., 2002), we finally conclude that BEWDS

[7] Note also that because every pair of contributions can be a possible pair of contributions under the MAM, the mechanism differs from the money back guarantee, assurance, or rebate rules mainly argued in the literature on the provision of a threshold public good (see Croson, 2008).

has a superior predictive performance compared with SPNE in our experiment.

In order to capture subjects' reasoning processes, we adopt three behavioral models that to a certain degree mimic BEWDS outcomes: *diagonalization* focuses on the diagonal line of the payoff table and considerably reduces computational effort in the MAM; the *minimax regret* criterion concerns foregone payoffs (Renou and Schlag, 2011); and *iterative best response* represents strategic sophistication (see Costa-Gomes and Crawford, 2006; Crawford and Iriberri, 2007). In the next step, in line with Cooper and Kagel (2005), we apply a coding scheme to quantify subjects' responses to open questions during and after the session. It turns out that 96.7% of subjects in the MAM were deemed to have found the diagonalization.[8] Moreover, 16.7% and 11.7% of subjects in the MAM and MCM respectively followed the subgame perfect minimax regret equilibrium strategy and approximately half of subjects in the MCM followed iterative best response.

The remainder of this paper is organized as follows. Section 2 overviews the MCM originally presented in Saijo et al. (2012). Section 3 introduces the MAM with examples and then presents the implementation results. Section 4 describes the proof of the uniqueness of the MAM. Section 5 presents SPNE predictions under the proposed mechanisms. Section 6 describes the experimental design. Section 7 discusses the experimental results, notably the superior predictive performance of BEWDS relative to SPNE. Section 8 offers three alternative models that mimic BEWDS and shows their validity through a coding scheme. Section 9 concludes.

2. The MCM

Consider a voluntary contribution game in the provision of a public good with two participants. Each player has an initial endowment $w > 0$ and he or she must decide the contribution $s_i \in [0, w]$. The sum of the contribution is multiplied by $\alpha \in (0.5, 1)$ and the benefit passes to every player, which expresses the non-rivalness of the public good. Then, $s_1 = s_2 = 0$ is the dominant strategy equilibrium and hence no public good is provided.

When there are only two levels of the contribution (i.e., 0 or w) and where players face the prisoner's dilemma, Saijo et al. (2012) introduced the notion of the MCM after the dilemma decision. By knowing the other player's choice between cooperation (contributing w and abbreviated by C) and defection (contributing zero and abbreviated by D) in the dilemma or the first stage, each player must approve (or y) or disapprove (or n) it. If both

[8] This observation illustrates that players are able to find a solution by instinct and experience rather than performing complex computations. We find another example in biochemical research on modeling the 3D structure of the protein that activates the AIDS virus. Although enormous simulation failed to provide a clear answer for decades, Khatib et al. (2011) sought ideas from the public through an online game called "Foldit", which give an improvement of model high score. Surprisingly, "Foldit" players, which had little background in the field, characterized the 3D structure in only three weeks.

players approve the choices of their counterparts in the first stage, the outcome or payoff vector is whatever they choose. If either player disapproves, however, the outcome is that when both choose *D*.

Figure 1 about here.

Figure 1 shows an example where w=10 and $\alpha = 0.7$. Note that this is an unusual extensive form game tree. Consider the case where player 1 chooses *D* and player 2 chooses *C* (termed subgame *DC* in Figure 1). If both choose *y* in the second stage, the payoff vector is (17,7). Otherwise, it is (10,10). Then, player 1 compares (17,10) with (10,10) and chooses *y* since *y* weakly dominates *n*. A vector *p* weakly dominates *q* if every component in *p* is at least greater than or equal to that in *q* and there is at least one component with strict inequality. Weak dominance between two choices is defined as weak dominance between their associated payoff vectors as usual. Similarly, player 2 chooses *n* since (10,10) weakly dominates (7,10). Therefore, the outcome in subgame *DC* is (10,10). In other words, (10,10) appears at the location of (*D*,*C*) of the reduced normal form game above the four subgames. By repeating the same procedure for every subgame, player 1 can construct the normal form game. Then, player 1 must compare (14,10) with (10,10). Since (14,10) weakly dominates (10,10), player 1 chooses *C*. Likewise, player 2 chooses *C* in the first stage and *y* in the second stage. This procedure describes BEWDS. Thus, BEWDS achieves (*C*,*C*) in the prisoner's dilemma game with the MCM.

Furthermore, neutrally stable strategies also attain (*C*,*C*), although the outcomes of NE and SPNE show every possibility, namely (*C*,*C*), (*C*,*D*), (*D*,*C*), and (*D*,*D*). As shown in Saijo et al. (2012), the strategies of BEWDS are a subset of neutrally stable strategies even though both achieve (*C*,*C*). The 19-round experiment with 10 randomly matched and unrepeated pairs presented in Saijo et al. (2012) showed almost full cooperation throughout the rounds. With the help of a related experiment, the authors also showed that as a behavioral principle BEWDS is most consistent with the data among the behavioral principles based upon NE, SPNE, neutrally stable strategies, or BEWDS. Following Saijo et al. (2012), we will use BEWDS as a basic behavioral principle, and restrict ourselves to the class of mechanisms where each player chooses $s_i \in [0, w]$ in the first stage and then *y* or *n* in the second stage. We call this mechanism an *approval mechanism*. In other words, the MCM is an approval mechanism.

However, the simple extension of the MCM[9] with two strategies 0 and w to the case with continuous strategies in $[0, w]$, namely if both players approve the contribution of the other in the first stage, then both contribute what they choose, whereas if either one disapproves, then both contribute zero, cannot implement the symmetric Pareto-efficient outcome. Consider the case with w=10 and $\alpha = 0.7$ again. For simplicity, each player can choose a contribution from {0,2,4,6,8,10}. Then, player 1's payoff table is shown in Table 1.

Table 1 about here.

The following two cases are examples of subgames. Consider subgame (0,6), the left-hand table in Table 2. Player 1 (the row player) chooses y since (14.2,10) weakly dominates (10,10) and player 2 chooses n since (10,10) weakly dominates (8.2,10).

Table 2 about here.

In other words, (10,10) is the payoff of this subgame. Similarly, both players choose y in subgame (6,10) where the payoff is (15.2, 11.2).

Table 3 shows player 1's payoff in the reduced normal form game. BEWDS deletes choices 0, 8, and 10 and hence choices 2, 4, and 6 survive in the first elimination. (14,14) Pareto dominates every payoff vector combined with choices 2, 4, and 6, and hence BEWDS1 cannot attain the symmetric Pareto-efficient outcome where 1 indicates the number of elimination rounds in the reduced normal form game and BEWDS without the number represents BEWDS1. Consider the next elimination round in which choice 6 is then eliminated. In the third elimination round, choice 4 is eliminated, and hence choice 2 is the final choice with BEWDS3. If the set of contributions is every integer between 0 and 10, (1,1) are the choices of BEWDSk for sufficiently large k. Since this example has finite possible contributions, there exists BEWDSk with k rounds; however, with continuous contributions, no BEWDS∞ exists, as the following proposition shows.

Table 3 about here.

---

[9] We call a MCM with many strategies MCM hereafter.

Proposition 1. *The MCM cannot implement the symmetric Pareto-efficient outcome under BEWDS and no BEWDS∞ strategy exists.*

Proof. Figure 2 shows player 1's payoff table of the reduced normal form game. Line 0-*l* shows that the payoff of player 1 is the same as the endowment. In other words, $u_1 = (w - s_1) + \alpha(s_1 + s_2) = w,$ and hence, $s_2 = ((1-\alpha)/\alpha)s_1.$ If the choice vector in the first stage is located in region *A*, player 1 chooses *n* in the second stage since the payoff with *y* is less than *w*, and hence player 1's payoff is *w* in the region.[10] Although player 1's payoff in region *B* is greater than *w*, his or her payoff ends up at *w* in the region since player 2 chooses *n*. Region $C = \{(s_1, s_2) : (1-\alpha)s_2 / \alpha \le s_1 \le \alpha s_2 / (1-\alpha)\}$ is the area where both players choose *y*. Since $\partial u_1 / \partial s_1 = -1 + \alpha < 0,$ fixing player 2's contribution and reducing player 1's contribution increase player 1's payoff. The arrow in Figure 2 shows this fact. Therefore, *w* is weakly dominated by $w - \varepsilon$ for sufficiently small $\varepsilon > 0$ for player 1 and hence $(s_1, s_2) = (w, w)$ is not a part of a BEWDS strategy profile. In other words, the MCM does not implement the symmetric Pareto-efficient outcome under BEWDS.

As Figure 2 shows, the choice $s_1 = ((1-\alpha)/\alpha)w$ (i.e., line *a*–*b*) weakly dominates any choice $s_1 \in [0] \cup ((1-\alpha)/\alpha)w, w]$. Furthermore, any choice $s_1 \in (0, ((1-\alpha)/\alpha)w]$ is not dominated by any other choices. Therefore, $s_1 \in (0, (1-\alpha)/\alpha)w]$ can go into the next round. Since square 0-*a*-*e*-*f* is similar to square 0-*c*-*d*-*g*, we can repeat the same procedure. Since $((1-\alpha)/\alpha) \in (0,1)$, $\lim_{k\to\infty}((1-\alpha)/\alpha)^k = 0,$ and hence no BEWDS∞ strategy exists. ∎

Figure 2 about here.

---

3. The MAM

In order to implement the symmetric Pareto-efficient outcome we introduce the MAM. In this mechanism, if both players approve the contribution of the other in the first stage, then both contribute what they choose, whereas if either one disapproves, then both contribute the minimum of the two contributions in the first stage. If this is the case, the player who chooses a larger number can obtain a refund.

Using Table 1, let us consider two examples of subgames under the MAM.

---

[10] We assume that players choose *y* when they are indifferent between *y* and *n* and that players eliminate all weakly dominated strategies at once whenever possible.

Consider first subgame (6,2). If either player disapproves the other's choice, the mechanism returns 4=6-2 to player 1, and hence players contribute (2,2) and the payoff is (10.8,10.8). Therefore, player 1 chooses *n* since (10.8,10.8) weakly dominates (9.6,10.8) and player 2 chooses *y* since (13.6,10.8) weakly dominates (10.8,10.8). In other words, (10.8,10.8) is the payoff outcome of this subgame.

Table 4 about here.

Next, consider subgame (8,10). Player 1 chooses *y* and player 2 chooses *n*; hence, the payoff is (13.2,13.2). In other words, player 1 approves (disapproves) player 2's first-stage choice if his or her first-stage choice is smaller (larger) than that of player 2. Table 5 shows player 1's payoff matrix in the reduced normal form game.

Clearly, choice 10 weakly dominates the rest, and hence (14,14) is the symmetric Pareto-efficient outcome under BEWDS. Table 5 presents a special structure: the payoff in subgame (*m,m*) is the same as that in (*m,n*) and (*n,m*) for all $n > m$, whereas the payoff in (*n,n*) is greater than that in (*m,m*) for all $n > m$. If the payoff matrix satisfies these two conditions, we say that it has an *echelon structure*.

Table 5 about here.

Proposition 2. *The MAM implements the symmetric Pareto-efficient outcome in BEWDS.*

Proof. First, we show that player 1's payoff matrix in the reduced normal form game under the MAM has an echelon structure. Fix $a \in [0, w]$, and choose any $b \in [0, w]$ with $b > a$. Consider subgame (*a,b*). Since $\{w - a + \alpha(a + a)\} - \{w - b + \alpha(a + b)\} = (b - a)(1 - \alpha) > 0$, player 2 chooses *n* in the second stage. Then, player 1's payoff becomes $w - a + \alpha(a + a)$. In other words, the payoff in subgame (*a,a*) is the same as that in (*a,b*) for all $b > a$ under the MAM. Similarly, we find that the payoff in subgame (*a,a*) is the same as that in (*b,a*) for all $b > a$ under the MAM. Choose any $a, b \in [0, w]$ with $b > a$. Then, the payoff in subgame (*b,b*) is greater than that in (*a,a*) since $w - b + 2\alpha b - (w - a + 2\alpha a) = (2\alpha - 1)(b - a) > 0$.

We further show that choice *w* weakly dominates $s_1 \in [0, w)$. Consider the case with $s_2 \in [0, s_1]$. From the first part of the echelon structure, we have $u_1(s_1, s_2) = u_1(s_2, s_2)$

$= u_1(w, s_2)$. [11] Consider the case with $s_2 \in (s_1, w)$. Using the two properties of the echelon structure, we have $u_1(s_1, s_2) = u_1(s_1, s_1) < u_1(s_2, s_2) = u_1(w, s_2)$. In other words, choice w weakly dominates $s_1 \in [0, w)$. ∎

It should be emphasized that Proposition 2 holds in more general environment. Even if two players have different preferences towards the public good $\alpha_1 \neq \alpha_2$, we can easily determine that the player with the higher (lower) first-stage choice still chooses n (y) under BEWDS. Further, if initial endowments are heterogeneous, say $w_1 \neq w_2$, the problem reduces to a symmetric case if each player chooses a ratio of contribution to his or her own endowment $s_i / w_i \in [0,1]$.

4. The MAM is Unique

There could be an infinite number of ways to implement the symmetric Pareto outcome using approval mechanisms under BEWDS. For example, consider the average approval mechanism (AAM): if either player disapproves, then both contribute the average of the two contributions in the first stage. Table 6 shows the payoff table for player 1 in the reduced normal form game under the AAM. This table shows that the full contribution of 10 units dominates every other contribution and thus that the symmetric Pareto outcome is attained by the AAM under BEWDS.

Moreover, under the AAM, the player who chooses a larger number can obtain a refund, whereas the player who chooses a smaller number must contribute more. In order to exclude this coercion we state that an approval mechanism is *voluntary* if each player's contribution does not exceed his or her first-stage choice. Clearly, the MAM satisfies voluntariness and the AAM does not. The budget-balanced mechanism proposed in Falkinger et al. (2000), under which subjects contributed over 90% of endowments, is similar to the AAM because it taxes (subsidizes) players with a contribution lower (higher) than the mean of the group.

Table 6 about here.

Moreover, we show that the MAM is a unique mechanism for satisfying

[11] Let $t = (t_1, t_2)$ be the decision in the second stage. Then, we must write $u_i((s_1, t_1), (s_2, t_2))$. Thus, $u_i(s_1, s_2)$ should be $u_i(s_1, s_2) = u_i((s_1, y), (s_2, n))$. In other words, we must specify the decision in the second stage in order to obtain the payoff of player 1. By slight abuse of notation, we can also use it both ways whenever the decision in the second stage is clear.

voluntariness andseveral other conditions. For this purpose, let us introduce the following definitions. What players choose in each stage $((s_1,t_1),(s_2,t_2))$ is referred to as a *path*. An approval mechanism g is a function that associates a contribution $g((s_1,t_1),(s_2,t_2))=(g_1((s_1,t_1),(s_2,t_2)),g_2((s_1,t_1),(s_2,t_2)))$ for every path $((s_1,t_1),(s_2,t_2))$. First, we require that approval mechanisms be *forthright* in the following manner: if both make the same first-stage choice or if both choose y in the second stage, then they contribute what they choose in the first stage; moreover, given first-stage choices, the outcomes when either chooses n are the same. Formally, we have:

Definition 1. An approval mechanism g is *forthright* if:
i) $g((s_1,\cdot),(s_2,\cdot))=(s_1,s_2)$ if $s_1=s_2$ where "$\cdot$" indicates either y or n;
ii) $g((s_1,y),(s_2,y))=(s_1,s_2)$ for every (s_1,s_2); and
iii) $g((s_1,n),(s_2,y))=g((s_1,y),(s_2,n))=g((s_1,n),(s_2,n))$ for every (s_1,s_2).

Note that forthrightness allows the possibility that the outcome when either player disapproves depends on their first-stage choices. In the following, vector inequality is taken component-wise as usual. The next property is voluntariness as argued above.

Definition 2. An approval mechanism g is *voluntary* if $g((s_1,t_1),(s_2,t_2))\leq(s_1,s_2)$ for every path $((s_1,t_1),(s_2,t_2))$.

The third property, *monotonicity*, is a plausible condition that reflects players' willingness to pay more. It states that if players weakly increase their first-stage choices, then their contributions weakly increase regardless of their decisions in the second stage.

Definition 3. An approval mechanism g is *monotonic* if $g((s_1',t_1),(s_2',t_2))\geq g((s_1,t_1),(s_2,t_2))$ for every (s_1',s_2') and (s_1,s_2) such that $(s_1',s_2')\geq(s_1,s_2)$ and every (t_1,t_2).

The above properties are sufficient for the uniqueness of the MAM.

Proposition 3. *Suppose that a forthright approval mechanism implements the Pareto-efficient allocation in BEWDS. If it is voluntary and monotonic, then it must be the MAM.*

Proof. Let g be a forthright approval mechanism that implements the Pareto-efficient allocation in BEWDS and that satisfies voluntariness and monotonicity. Let (s_1,s_2) be any first-stage choice. If $s_1=s_2$, $g((s_1,\cdot),(s_2,\cdot))=(s_1,s_2)$ according to condition i) of

forthrightness, which shows the outcome of the MAM. Consider the case with $s_1 \neq s_2$. Without loss of generality, assume that $s_1 > s_2$ and consider two cases: $s_1 = w$ and $s_1 < w$. First, consider subgame (w, s_2) with $w > s_2$. In order to implement the efficient allocation in BEWDS, w must weakly dominate s_2 for player 2 in the reduced normal form game. In other words,

(1) $u_2^g((w,\cdot),(w,\cdot)) \geq u_2^g((w,t_1),(s_2,t_2))$

where u_i^g is i's payoff function induced by g[12] and t_i is player i's decision under BEWDS.

Suppose that $t_1 = t_2 = y$. Then, according to conditions i) and ii) of forthrightness, we have

(2) $u_2^g((w,y),(s_2,y)) = (w - s_2) + \alpha(w + s_2) > (w - w) + \alpha(w + w) = u_i^g((w,\cdot),(w,\cdot))$

which contradicts (1). Thus, $t_1 = n$ or $t_2 = n$. Together with condition iii) of forthrightness, (1) and (2) show that $u_2^g((w,y),(s_2,y)) > u_2^g((w,y),(s_2,n)) = u_2^g((w,n),(s_2,y)) = u_2^g((w,n),(s_2,n))$. Hence, y weakly dominates n for player 2 in the second stage, namely $t_2 = y$, which implies $t_1 = n$.

Consider now the reduced normal form game under g and compare the outcomes in subgames (s_2, s_2) and (w, s_2). Since player 1 chooses w by assumption, we have $u_1^g((w,n),(s_2,y)) \geq u_1^g((s_2,\cdot),(s_2,\cdot))$. Let $g((w,n),(s_2,y)) = (g_1, g_2)$. Then, $u_1^g((w,n),(s_2,y)) = w - g_1 + \alpha(g_1 + g_2) \geq w - s_2 + \alpha(s_2 + s_2) = u_1^g((s_2,\cdot),(s_2,\cdot))$. Since g is voluntary, $(g_1, g_2) \leq (w, s_2)$. Because of the monotonicity of g, $(g_1, g_2) \geq (s_1, s_2)$. The intersection of these three inequalities shows $(g_1, g_2) = (s_2, s_2)$. Thus, according to condition iii) of forthrightness, we have

(3) $g((w,n),(s_2,y)) = g((w,y),(s_2,n)) = g((w,n),(s_2,n)) = (s_2, s_2)$.

In other words, both players contribute s_2, which is the minimum of (w, s_2) when at least one player chooses n.

Now take any $s_1 \in (s_2, w)$. Together with $(w, s_2) \geq (s_1, s_2) \geq (s_2, s_2)$ and (3), since g is monotonic, we have $(s_2, s_2) = g((w,n),(s_2,y)) \geq g((s_1,n),(s_2,y)) \geq g((s_2,n),(s_2,y)) = (s_2, s_2)$. Hence, we have $g((s_1,n),(s_2,y)) = (s_2, s_2)$. In other words, g is the MAM. ∎

As far as a forthright approval mechanism implements the Pareto-efficient

[12] We omit superscript g when a mechanism is specified.

allocation in BEWDS, Proposition 3 states that the mechanism is voluntary and monotonic if and only if it is the MAM. Although it is artificial, the following forthright mechanism is voluntary, but not monotonic. Consider a forthright rule in which a player who announces a bigger contribution than the other does in the first stage must contribute zero and the smaller player must contribute what he or she chooses in the first stage if either one chooses n in the second stage. Apparently, this mechanism is voluntary. However, it is not monotonic since $g_1((6,n),(2,y)) = 0 < g_1((2,n),(2,y)) = 2$. [13]

5. SPNE outcomes under the MCM and MAM

A subgame perfect Nash equilibrium is often used to analyze multi-stage games. We first show that under the MCM, any pair of contributions in region C in Figure 2 is supported by SPNE.[14] Although the MAM under BEWDS chooses a Pareto-efficient outcome d in Figure 2, any symmetric pair of contributions is supported by SPNE. Fix an approval mechanism g. Player i's strategy is a pair of a first-stage choice s_i and a function $t_i(\cdot)$ that associates y or n to each first-stage choice (s_1, s_2). A strategy profile $(s_1, t_1(\cdot)),(s_2, t_2(\cdot))$ of the game induced by g is a Nash equilibrium if for each i, j with $j \neq i$, $u_i^g((s_i, t_i(\cdot)),(s_j, t_j(\cdot))) \geq u_i^g((s_i', t_i'(\cdot)),(s_j, t_j(\cdot)))$ for all $(s_i', t_i'(\cdot))$. A Nash equilibrium $((s_i, t_i(\cdot)),(s_j, t_j(\cdot)))$ is subgame perfect if $((s_i, t_i(\cdot)),(s_j, t_j(\cdot)))$ constitutes a Nash equilibrium in every subgame.

To illustrate an example of SPNE, consider the case with w=10 and $\alpha = 0.7$ again, and let a strategy profile be ((6,t(·)), (8,t(·))), where $t(6,8) = y$ and $t(s_1, s_2) = n$ if $(s_1, s_2) \neq (6,8)$. The left-hand side of Table 7 shows the payoff matrix for subgame (6,8). The shaded cells indicate the Nash equilibrium outcomes of the subgame. Then,a Nash equilibrium of subgame (6,8) is (y,y), which yields (13.8,11.8). Moreover, in every subgame, (n,n) is another Nash equilibrium since the outcomes of (n,y), (n,y), and (n,n) are the same. Thus, the strategy profile ((6,t(·)), (8,t(·))) constitutes a Nash equilibrium in every second-stage subgame.

Table 7 about here.

Thus, both players get 10 unless they choose (6,8) and hence (6,8) is a Nash

[13] It is easy to check that the mechanism implements a Pareto-efficient allocation in BEWDS.

[14] Given an equilibrium concept, we say contributions (q_1,q_2) are supported by an equilibrium if there exists an equilibrium where contributions are (q_1,q_2).

equilibrium of the reduced normal form game. Therefore, the strategy profile is a subgame perfect Nash equilibrium.

The drawback of the above subgame perfect Nash equilibrium is that players play the Pareto-dominated Nash equilibrium off the equilibrium path in region *C* in Figure 2. For the subgame (4,8) shown on the right-hand side of Table 7, we see that a Nash equilibrium (*n*,*n*) with payoffs (10,10) is weakly dominated by (*y*,*y*) with payoffs (14.4,10.4). BEWDS eliminates such possibilities since players approve if the first-stage choices are located in region *C* including both (6,8) and (4,8). In a similar way, all contributions in region *C* can be supported by SPNE. To summarize, we have Proposition 4.

Proposition 4. *Under the MCM, i) for any first-stage choices* (s_1, s_2)*, there exists a subgame perfect Nash equilibrium where players choose* (s_1, s_2)*; further, ii) contributions* (q_1, q_2) *are supported by SPNE if and only if* $(1-\alpha)q_2 / \alpha \le q_1 \le \alpha q_2 / (1-\alpha)$.

Proof. See Appendix. ∎

By contrast, the set of SPNE contributions under the MAM coincides with the set of symmetric contribution profiles. In contrast to the MCM, no subgame perfect Nash equilibrium strategy involves the Pareto-dominated Nash equilibrium.

Proposition 5. *Under the MAM, i) for any symmetric first-stage choices* (s,s)*, there exists a subgame perfect Nash equilibrium where players choose* (s,s) *and contribute* $(q_1, q_2) = (s,s)$*; further, ii) there exists no subgame perfect Nash equilibrium with asymmetric first-stage choices or contributions.*

Proof. See Appendix. ∎

From Propositions 2, 3, 4, and 5, we see that both in the MAM and MCM the set of BEWDS paths is a subset of SPNE paths, although the former is not a refinement of the latter.

6. Experimental Design

We conducted four treatments in order to test the performances of the approval mechanisms. The VCM was the control treatment, while the other three treatments were the MCM, MAM, and SMAM. The SMAM also implements the efficient allocation in BEWDS. To consider the SMAM, we relax the assumption that both players proceed to the

second stage. Recall that under the MAM, the player who follows BEWDS approves (disapproves) if his or her choice is smaller (larger) than that of the other player and is indifferent between approval and disapproval when both players make the same choice (see Section 3). Hence, the decision by the player with the higher first-stage choice alone determines the outcome in every second stage. Then, we define the SMAM as follows. If both players make the same first-stage choice, the game ends. Otherwise, only the player with the higher first-stage choice proceeds to the second stage. If he or she approves, they contribute what they choose in the first stage. If he or she disapproves, both players contribute the minimum of their choices in the first stage.[15]

Each of the above-mentioned four treatments had three experimental sessions. These 12 sessions were all conducted at Osaka University in March, April, September and November 2011. Subjects were recruited from Osaka University through campus-wide advertisement and were inexperienced in this particular type of experiment. No individual participated in more than one session. The experiment was computerized using the experimental software z-Tree (Fischbacher, 2007). In each session, 20 subjects participated. Each was seated at a computer terminal assigned by lottery. All terminals were separated by partitions. No communication between subjects was allowed.

Each subject had a set of printed instructions, a record sheet, and a payoff table (these materials are included in Appendix D; the payoff table was called the points table in the experiment). Table A8 shows the payoff table for every treatment. Each subject was considered to be player 1. The first column of this table lists player 1's contribution to the public good (called "your investment"), while the first row lists player 2's contribution to the public good (called "the investment of your counterpart"). Each cell contains both players' payoffs, shown in blue and red for player 1 and 2, respectively. Player i's payoff is given by $u_i = 300\{(24 - q_i) + 0.7(q_1 + q_2)\}$ for each contribution (q_1, q_2). Since we consider two players with identical utility functions and endowments, the payoff table is identical for all subjects. Subjects were informed that they have identical payoff tables.

Instructions were read aloud by an experimenter for approximately 10 minutes. After that, subjects were given another 10 minutes to ask questions. Then, we proceeded to the payment periods. There was no practice period. Each session consisted of 19 periods under the perfect stranger matching protocol. We informed subjects that each of them would meet any other subject in each period.

The MAM and MCM treatments continued as follows. At the beginning of each period, all subjects endowed with 24 tokens were anonymously matched into pairs. In the first stage (called the "choice stage"), subjects were asked to enter their contributions as

[15] The SPE outcome of the SMAM is the same as that of the MAM.

nonnegative integers into a box in the display and write down their choices along with their reasoning in the corresponding row of the record sheet. Once all subjects had finished their tasks, they clicked the OK button.

In the second stage (called the "decision stage"), the first-stage choices of both players and the payoff matrix in the second stage were displayed. After all players wrote down the first-stage choices of their counterparts, they chose to approve or disapprove by clicking the radio buttons and recording the decision along with the reasoning. Once all subjects in the second stage had finished the above tasks and clicked the OK button, they proceeded to the results screen.

The results screen included the first-stage choices of both players, their "approve" or "disapprove" decisions, and their payoffs (points earned) in the period. No information on the choices of the other nine groups was provided to subjects. Finally, subjects wrote down the decisions of their counterparts and the points they earned and clicked the Next button to begin the next period.

If a subject in the SMAM did not proceed to the second stage, he or she simply wrote down the first-stage choice of his or her counterpart before other subjects finished the second stage and circled "none" while facing the waiting screen. The treatment without the decision stage became the VCM treatment.

After playing the 19 periods, subjects completed a questionnaire and they were paid privately in cash immediately. Each subject was paid an amount proportional to the sum of the points that he or she had earned for the 19 periods. Individual payments ranged from \$49.91 to \$87.97.

7. Experimental Results

7.1. Average contributions

Figure 3 shows the time path of average contributions over the 19 periods arranged by mechanism. Individual contributions were evaluated after the second stage, except for the baseline VCM. Table 8 compares contributions for the four treatments using a two-tailed Mann–Whitney test.

Figure 3 about here.

We first assess the results of the VCM. The time path of average contribution in the VCM shows a similar pattern to that of previous VCM experiments (e.g., Ledyard,

1995). In the first round, subjects contributed 15.3% of endowments (3.68 tokens), with contribution rates gradually decreasing to 5.7% (1.37 tokens) in the last period. The hypothesis test based on Spearman's rank correlation coefficient showed that the downward trend in average contributions was statistically significant ($\rho = -0.9158, p = 0.00005$). When the data were pooled across all 19 periods and three sessions, subjects contributed 10.2% of endowments (2.44 tokens). Out of 570 outcomes (10 pairs × 19 periods × 3 sessions), there was no case where both players contributed all endowments.

Table 8 about here.

Result 1. *Subjects in the MAM successfully sustained cooperation and contributed on average 94.9% of endowments when pooled across all 19 periods and three sessions.*

Introducing the MAM facilitates almost full contributions among subjects, even in the earlier period. Contributions in the first period averaged 76.9% of endowments (18.45 tokens). They then rose repeatedly for the first five periods, achieving 97.8% (23.47 tokens) in period 5. Then, cooperation was sustained with an average contribution rate of over 95% in every period except for the final one. Moreover, the convergence to the efficient allocation under the MAM was statistically supported by Spearman's rank correlation test to examine the upward trend in average contributions ($\rho = 0.5566, p = 0.009$). When the data were pooled across all 19 periods and three sessions, subjects in the MAM contributed on average 94.9% of endowments (22.78 tokens). Out of 570 outcomes, there were 475 outcomes where both players contributed all endowments. A two-tailed Mann–Whitney test showed that subjects in the MAM contributed significantly more than those in the VCM with the test statistic z=9.445 (p<0.001).[16]

Result 2. *The simplification of the second stage significantly decreased the average contribution rate in the SMAM compared with the MAM.*

The contributions in the SMAM show similar patterns to those in the MAM. It

[16] The null hypothesis is that contribution rates are the same between the MAM and VCM. We used the same method as that presented by Andreoni and Miller (1993) and Charness et al. (2007). We first calculated the average contribution rate of each subject across periods and then calculated the test statistic using these averages in order to eliminate cross-period correlation.

took five periods to achieve a contribution rate of over 90% and almost all subjects maintained full contributions until the end of the session. Spearman's rank correlation test shows that the upward trend in average contributions was statistically significant $(\rho = 0.62484, p = 0.004)$. Overall, the average contribution rate of the SMAM was 89.9%. Out of 570 outcomes, there were 424 outcomes where both players contributed all endowments. The contribution rate was also significantly higher than that of the VCM (two-tailed Mann–Whitney test, z=9.275, p<0.001). However, the contribution rate of the SMAM was significantly lower than that of the MAM (two-tailed Mann–Whitney test, z=3.799, p<0.001).

Result 3. *The MCM increases contributions relative to the VCM, but it does not lead to the efficient provision of the public good.*

In contrast to the MAM and *S*MAM, only seven out of 540 pairs of the MCM achieved full contributions and the approval of both players. The average contribution under the MCM fluctuated within a relatively narrow range, but decreased repeatedly, and remained higher than that of the VCM in every period. Spearman's rank correlation test showed that the downward trend in average contributions was statistically significant. The overall average contribution rate of the MCM was 28.4% (6.83 tokens), which was significantly higher than that of the VCM (two-tailed Mann–Whitney test, z= 7.284, p<0.001). However, the contribution rate was significantly lower than that of the MAM and SMAM (two-tailed Mann–Whitney test, z=9.450 and p<0.001 for the test of the MCM vs. MAM and z=9.133 and p<0.001 for MCM vs. SMAM).

7.2. *Why was the contribution rate in the SMAM significantly lower than that in the MAM?*

We argue that the significant difference shown in Result 2 occurred mainly because of the choices of two subjects in the third session of the SMAM.

Result 4. *A simplification of the second stage does not have a significant effect on the average contribution rate when we compare the MAM with the SMAM after excluding the data from groups including either of the two subjects who did not understand the rules of the experiment until the end and who strategically kept making low first-stage choices.*

Table 9 shows average contribution rates per subject. In the MAM, all subjects contributed on average more than 80% of endowments. In the SMAM, by contrast, one subject (subject 13 in the third session) contributed only 5.26% of endowments (24 tokens in

total across 19 periods) throughout the session.

Table 9 about here.

According to the record sheet and post-experiment questionnaire, this subject seems to have thought that he or she has to allocate 24 tokens across 19 periods even though he or she wanted to contribute more.[17] Clearly, this subject did not understand the rules of the experiment. Furthermore, all of his or her paired subjects except for one chose disapproval because it was beneficial for them. This resulted in a lower contribution rate in the SMAM compared with that in the MAM. If we exclude the data from subject 13 in the third session and those from the other subjects when they were matched with him or her, the average contribution rate of the SMAM rose to 92.3%, still significantly lower than that of the MAM at the 5% significance level (two-tailed Mann–Whitney test, z=-2.395, p=0.017). Subject 9 in the third session also contributed far less compared with the other subjects. He or she contributed on average 57.5% of endowments. This subject intentionally chose low contributions, expecting that his or her matched subjects might approve by giving up a small amount of their points and allowing him or her to earn more points. However, his or her contributions were disapproved by all 17 counterparts, who proceeded to the second stage.[18] Similar to the above case, we also performed a two-tailed Mann–Whitney test after excluding the data from the groups in which subjects 9 or 13 in the third session participated. Then, the contribution rate of the SMAM became 94.0%, which was not significantly different from that of the MAM at the 10% level (z=-1.392, p=0.164).

7.3. Evaluation of the predictive performances of BEWDS and SPNE

7.3.1. Evaluation based on the equilibria of the whole game

From the aggregate data of the MAM and SMAM, subjects' choices are apparently consistent with BEWDS since they sustainably contributed nearly 95% of their endowments. In this subsection, we thus examine the data from individual groups in order to provide more evidence that BEWDS better describes the data compared with SPNE.

In order to evaluate several equilibrium concepts and learning models, the

---

[17] In the post-experiment questionnaire, this subject wrote: "Many subjects chose 24. … I could not contribute more because I have only 24 tokens."

[18] In the post-experiment questionnaire, this subject wrote: "My matched subjects chose a higher contribution than I did, and so my choice was disapproved in most periods. However, I wanted to choose the number, expecting that my matched subjects might approve."

behavioral game theory literature sometimes abstracts the economic environment and sets various types of games in which each equilibrium concept or learning rule predicts a unique outcome and their predictions differ. Thus, performance can be evaluated based on how close the experimental data and these equilibrium predictions are, for example by analyzing mean squared deviations (see Ert et al., 2011).[19] By contrast, in our study the public good environment and mechanisms are given. Thus, the BEWDS paths are included in the SPNE paths and may not always be unique as shown in the theoretical section. These facts lead to two problems. One problem is how to evaluate the inclusion of equilibrium paths. Another is overcoming the difficulty of defining the distance between the experimental data and equilibrium predictions because of their multiplicity. Hence, we need to evaluate the equilibrium in an alternative manner.

To solve these problems, we use two criteria to evaluate BEWDS and SPNE for each approval mechanism treatment g.[20] The first criterion is the number of equilibrium paths, while the second is the proportion of the path data $((s_1,t_1),(s_2,t_2))$ that is consistent with the equilibrium paths (for the derivation of the equilibrium paths including BEWDSk, k=2,3,4,5, see Appendix B). We conclude that one equilibrium concept has a superior predictive performance compared with the other concept under g if this (first) concept has fewer paths than the other concept and the proportion of equilibrium-consistent path data is significantly higher than the other. The followings are for the case of a tie in one criterion. If the two equilibrium concepts have an equal number of equilibrium paths and the proportion of the consistent path data of one is significantly higher than the other, this (first) concept shows a superior performance. Further, if one equilibrium concept has fewer paths than the other and the proportion of the consistent path data is not significantly different, this (first) concept shows a superior performance.

The statistical significance of the difference in the proportions of equilibrium-consistent paths between BEWDS and SPNE, denoted by b^g and p^g, respectively, is tested using McNemar's exact test for the null hypothesis $H_0 : b^g = p^g$ against the alternative $H_1 : b^g \neq p^g$.

Note that these two criteria are complementary. If our evaluation were based only on the proportion of equilibrium-consistent path data, then the trivial equilibrium concept that predicts all possible paths would not be bettered by any other equilibrium concept.

---

[19] Using games with a unique equilibrium has another advantage, namely avoiding multiple supergame equilibria. Using various types of games (e.g., entry games and trust games) helps confirm the robustness of equilibrium or learning models (see Erev and Roth (1998) and Nyarko and Schotter (2002) for strategic form games and Stahl and Haruvy (2009) and Ert et al. (2011) for extensive form games). By contrast, our situation is similar to the experimental studies of Brandts and Holt (1992), Banks et al. (1994), and Cooper and Kagel (2008), which evaluated Nash refinements using signaling games.

[20] We omit data on the VCM since both BEWDS and SPE predict zero contributions.

Thus, we can avoid this problem by preferring a smaller number of equilibrium paths. By contrast, even if the number of equilibrium paths were small, it would not serve our purpose to find subjects' behavioral principles as long as the equilibrium paths failed to fit the data. Thus, the proportion of equilibrium-consistent path data should always be considered. Therefore, we have:

Result 5. *i) In the MAM, the number of BEWDS paths is four, while that of SPNE paths is 100. There is no significant difference in the proportion of equilibrium-consistent path data between BEWDS (83.3%) and SPNE (83.9%).*
ii) In the SMAM, the number of BEWDS paths is one, while that of SPNE paths is 25. There is no significant difference in the proportion of equilibrium-consistent path data between BEWDS (74.4%) and SPNE (74.6%).
iii) In the MCM, the number of BEWDS paths is 123, while that of SPNE paths is 1160. The proportion of BEWDS-consistent path data (46.3%) is significantly smaller than that of SPNE (86.8%).

Table 10 lends support to Result 5. The third column shows the equilibrium paths. Let us begin with the MAM.

Table 10 about here.

First, as shown in the fourth column of Table 10, BEWDS predicts four payoff-equivalent paths ((24,·),(24,·)), while SPNE predicts 100 paths where players make the same first-stage choice. Dividing the number of equilibrium-consistent paths in the fifth column by the total number of paths suggests that BEWDS explains 83.3% (=475/570) of the data in the MAM, slightly lower than SPNE (83.9% =478/570). However, McNemar's exact test does not reject H_0 at the 10% significance level. Since BEWDS has a smaller number of equilibrium paths than SPNE and because the proportions of the equilibrium-consistent data are not statistically different between them, we conclude that BEWDS is more predictive than SPNE in the MAM.

We obtain the same conclusion in the SMAM, where BEWDS predicts (24,24) and SPNE 25 paths. SPNE can thus explain only one additional data point that BEWDS cannot (74.4% =424/570 for BEWDS vs. 74.6% =425/570 for SPNE). Again, McNemar's exact test does not reject H_0 at the 10% significance level.

We obtain inconclusive results in the MCM, however, in which BEWDS predicts 123 paths and SPNE 1160 paths. Further, SPNE explains nearly twice as many observed paths as BEWDS (86.8% =495/570 vs. 46.3% =264/570). McNemar's exact test rejects H_0 at the 1% significance level, suggesting that SPNE explains a significantly larger proportion of data than BEWDS. Hence, we cannot determine whether BEWDS or SPNE has a superior predictive performance.

7.3.2. Evaluation based on the equilibria of subgames

Thus far, we have only examined whether path data are consistent with some equilibrium path. In this subsection, we focus on second-stage decisions. By fixing mechanism treatment *g* and equilibrium concept BEWDS (SPNE), we say that the path data $((s_1,t_1),(s_2,t_2))$ is *subgame-consistent* if second-stage decisions (t_1,t_2) are consistent with the elimination of weakly dominated strategies for both players (NE) in subgame (s_1,s_2) under *g*.[21] Similar to the argument in the previous subsection, we perform McNemar's exact test for the equality of the proportion of subgame-consistent path data between BEWDS and SPNE by treatment. We omit the path data where both subjects are indifferent between approval and disapproval. Then, we have:

Result 6. *i) In the MAM, BEWDS prediction is unique, while there are two SPNE predictions in every subgame. There is no significant difference in the proportion of subgame-consistent path data between BEWDS (85.9%) and SPNE (90.2%).*
ii) In the SMAM, both BEWDS and SPNE predict a unique and identical decision in every subgame. The proportion of subgame-consistent path data is 97.2%.
iii) In the MCM, BEWDS predicts one fewer decision than SPNE in every subgame. There is no significant difference in the proportion of subgame-consistent path data between BEWDS (86.3%) and SPNE (86.8%).
iv) Among the proportions of second-stage decisions consistent with neither BEWDS nor SPNE, (n,y) when $s_1<0.7s_2/0.3$ in the MCM ranks at the top with 15.8%, followed by (y,y) in the MAM with 9.8%, (y,y) in the MCM when $s_1>0.7s_2/0.3$ with 3.4% and (y,-) in the SMAM with 2.8%.

Table 11 lists the frequencies of subgame-consistent path data under BEWDS and SPNE by treatment. In the MCM, second-stage subgames are classified into three rows

[21] In order to deal with multiple equilibrium concepts and multiple equilibria in the imperfect information games induced by our mechanisms, we slightly modified the definition of subgame consistency for perfect information game experiments proposed by Binmore et al. (2002). This requires that for every subgame, if subjects reach there, they make the choices prescribed by the unique subgame perfect Nash equilibrium. For perfect information games, our subgame consistency is identical to that presented by Binmore et al. (2002).

depending on BEWDS predictions. The middle four columns list the frequencies of the second-stage decisions. The shaded cells represent subgame consistency under BEWDS. Tables A9 and A10 show the BEWDS-consistent decisions for each second-stage subgame. As shown in the proofs of Propositions 4 and 5, SPNE predictions contains (n,n) other than BEWDS predictions in every subgame of both the MAM and MCM. Note that both BEWDS and SPNE predict n in every subgame of the SMAM.

A large proportion of path data is subgame-consistent under BEWDS for every approval mechanism treatment, as shown in the third column from the right. The proportions are 85.9% (=79/(9+79+4)) in the MAM, 97.2% (=141/(4+141)) in the SMAM, and 86.3% (=(377+2+2+111)/570) in the MCM.

Table 11 about here.

For SPNE, by contrast, these proportions are 90.2% (=(79+4)/(9+79+4)) in the MAM, 97.2% (=141/(4+141)) in the SMAM, and 86.8% (=(377+2+2+111+1+2)/570) in the MCM. McNemar's exact test does not reject the equality of the proportion of subgame-consistent path data between BEWDS and SPNE at the 10% significance level in all treatments.[22] In other words, SPNE does not significantly improve the explanation of our data compared with BEWDS. By applying the evaluation presented in the previous subsection, we conclude that BEWDS is more predictive than SPNE in terms of subgame consistency both in the MAM and in the MCM.

Table 10 reports that the number of SPNE paths is 1160. Note that many of these SPNE paths are constructed so that subjects play a Pareto-dominated Nash equilibrium (n,n) on off-the-equilibrium paths at $s_1<0.7s_2/0.3$, as shown in Section 5. However, as we observed only one case of (n,n), it is hard to justify that subjects follow SPNE.

It is uncertain why the violation of both BEWDS and SPNE occurs with a high frequently of 15.8% =71/(377+71+1) in the MCM as (n,y) when $s_1<0.7s_2/0.3$ compared with other cases: (y,y) in the MAM with 9.8% =9/(9+79+4); $(y,\text{-})$ in the SMAM with 2.8% =4/(4+141)); and (y,y) in the MCM when $s_1>0.7s_2/0.3$ with 3.4% =4/(4+111+2)). Note that when $s_1<0.7s_2/0.3$, player 1 can be better off compared with his or her initial endowment when both players approve. Hence, if player 1 deliberately chooses n while knowing that player 2 chooses y following BEWDS, player 1 loses his or her payoff.

[22] Since both BEWDS and SPE predict n in the SMAM, it is natural that McNemar's exact test provides a p-value of 1.000.

Inequality aversion explains these observations. Consider subgame (11,5) in the MCM using Table A8. Because 11>5, player 2, who follows BEWDS, chooses *y*. Then, if player 1 chooses *y*, he or she will be behind player 2 by 9060-7260=1800. If player 1 chooses *n*, both players obtain the same payoff of 7200 and he or she is not disadvantaged in terms of inequality. Hence, player 1 will choose *n*. A similar argument shows that inequality-averse player 1 disapproves player 2's choice when $s_1 > s_2$ in general. The questionnaire analysis discussed in the next section shows that subjects who choose weakly dominated disapproval are deemed to follow inequality aversion.

Although inequality aversion seems to be anomalous behavior in the MCM, it provides new insights into the effectiveness of the MAM. Inequality-averse player 1 still disapproves player 2's choice when $s_1 > s_2$. This is consistent with BEWDS. In other words, even when a pair consists of a follower of inequality aversion and that of BEWDS, players behave in the same manner in the second stage of the MAM. Although we do not formulate the formal model for inequality aversion here, this fact suggests that MAM directs players with heterogeneous behavioral rules towards cooperation. In the next section, we explore this possibility in greater depth as a step towards designing workable public good mechanisms.

8. Reasoning Processes in the MAM and MCM

As shown in Section 7, BEWDS has more predictive power in terms of path data than SPNE both in the MAM and in the SMAM. However, this result does not imply that subjects follow the underlying logic of BEWDS in the MAM. Subjects need high computational ability to eliminate weakly dominated strategies in every subgame, especially in the reduced normal form game that has a 25×25 payoff table.

In this section, we investigate what underlying reasoning processes subjects follow by examining their responses to the open-ended questions during and after the experiment. First, we propose alternative models whose predictions are consistent with the data in the MAM and MCM, according to the notable descriptions in the record sheets completed during the experiment and to the post-experiment questionnaire. Then, we employ a coding method in order to count the number of decisions that mention the idea of alternative models. Furthermore, we count the number of decisions that mention the idea of inequality aversion in the record sheet when players chose *n* even though BEWDS prediction was *y* in order to support the conjecture presented in subsection 7.3.2.

8.1. Reasoning Processes in the MAM

Figure 4 illustrates the distribution of first-stage choices in the MAM. The striking

feature of the data is that most subjects choose the full contribution. We now propose two alternative models that predict that players make the full contribution (or the symmetric Pareto-efficient outcome).

Figure 4 about here.

8.1.1. Diagonalization

The first alternative model simplifies backward induction. Some subjects seemed to understand that if players in a pair chose different first-stage choices, then the player with the higher contribution would disapprove the other's choice. This results in the same contribution for both players. Therefore, such players focused on the payoffs on the diagonal line of the payoff table. Given this expectation, subjects found that the full contribution maximizes their own payoffs among the contributions on the diagonal line. We call this heuristic or algorithm *diagonalization*, which can be divided into two parts:

(*D*-1) Since a player who chooses the higher first-stage choice will disapprove the other's choice, both obtain the payoffs of the diagonal line in the payoff table; and
(*D*-2) Full contributions for both players attain the maximum payoff among the payoffs on the diagonal line.

Then, we have the following simple proposition.

Proposition 6. *If both players follow diagonalization in the MAM, then the outcome is symmetric Pareto-efficient.*

8.1.2. Regret minimization

The second alternative model is *regret minimization* coupled with subgame perfection. Some subjects imagine foregone payoffs from unchosen strategies. This idea is close to the simplest version of ε-minimax regret equilibrium introduced by Renou and Schlag (2011), who incorporated into game-theoretic models experimental evidence that people avoid ambiguity and the decision-theoretic formulation of such behavior. Let g be an approval mechanism. For every pure strategy profile $((s_1, t_1(\cdot)),(s_2, t_2(\cdot)))$, player i's *regret* at $((s_1, t_1(\cdot)),(s_2, t_2(\cdot)))$ is defined by $R_i((s_i, t_i(\cdot)),(s_j, t_j(\cdot))) = \max_{(s_i', t_i'(\cdot))} u_i^g((s_i', t_i'(\cdot)),(s_j, t_j(\cdot))) - u_i^g((s_i, t_i(\cdot)),(s_j, t_j(\cdot)))$, $j \neq i$. This is the

difference between his or her maximal payoff when player *i* best responds to *j*'s strategy $((s_j, t_j(\cdot))$ and his or her payoff when he or she plays $((s_i, t_i(\cdot))$. A strategy profile $((s_1, t_1(\cdot)),(s_2, t_2(\cdot)))$ is a *minimax regret equilibrium* if for each *i*, *j* with $j \neq i$, and for all $(s_i', t_i'(\cdot))$, $\max_{(s_j', t_j'(\cdot))} R_i((s_i, t_i(\cdot)),(s_j', t_j'(\cdot))) \leq \max_{(s_j', t_j'(\cdot))} R_i((s_i', t_i'(\cdot)),(s_j', t_j'(\cdot)))$. In other words, each player chooses the strategy that minimizes the largest possible regret in the equilibrium.[23] We say that a minimax regret equilibrium is *subgame perfect* if the action profile induced by $((s_1, t_1(\cdot)),(s_2, t_2(\cdot)))$ constitutes a minimax regret equilibrium in every subgame.

Proposition 7. *The MAM implements the symmetric Pareto-efficient outcome in the subgame perfect minimax regret equilibria.*

Proof. See Appendix. ■

Propositions 2 and 7 together imply the double implementation result.

Corollary. *The MAM doubly implements the symmetric Pareto-efficient outcome in BEWDS and the subgame perfect minimax regret equilibria.*

8.2. Reasoning Processes in the MCM

As shown in subsection 7.3, BEWDS does not satisfactorily explain the data in the MCM. Figure 5 illustrates the distribution of first-stage choices in the MCM. Remember that in the MCM BEWDS predicts any contribution between one and 11 tokens. Figure 5 shows that 805 out of 1140 first-stage choices (70.6%) are in this range. However, BEWDS alone cannot explain the additional feature of the distribution that the spikes appear at five, six, 11, and 24 tokens. This indicates that some reasoning processes other than BEWDS might work. The alternative models presented here partially capture this feature.

Figure 5 about here.

8.2.1. Regret minimization

Certain subjects in the MCM also described a reasoning process that is consistent

[23] Here, we assume *complete uncertainty*, namely each player believes that his or her opponent may choose any strategy.

with regret minimization.

Proposition 8. *A unique contribution supported by the subgame perfect minimax regret equilibria coincides with the maximal one supported by BEWDS under the MCM.*

Proof. See Appendix. ■

Although regret minimization can explain the highest spike (11 tokens) in Figure 6, it does not predict the spikes at five, six, and 24 tokens.

8.2.2. Iterated Best Response

An alternative model for the MCM is *iterated best response* in the reduced normal form game coupled with the elimination of weakly dominated strategies in the second stage.[24] Selected subjects seemed to expect that the other player would choose 24 tokens in the first stage. Assume that player 1 expects player 2 to choose 24 tokens and to eliminate weakly dominated strategies in the second stage. As shown in Table A9, player 1 will know that player 2 disapproves if and only if player 1 chooses no more than 10 tokens. Then, player 1 can easily see that contributing 11 tokens is the best response. Another step of iterated reasoning leads player 1 to contribute five tokens, which is the best response to the 11 tokens of player 2.[25] To put this procedure formally, let $B(k)$ be a sequence such that $B(0) = w$ and define iteratively $B(k) = BR(B(k\text{-}1))$ until $B(k) = B(k\text{-}1)$, where $BR(.)$ is the best response to $B(k\text{-}1)$ in the reduced normal form game. By contrast, the choice that survives BEWDSk is $\{1, 2,\ldots, B(k)\}$. Then, we have the following result.

Proposition 9. *A sequence of first-stage choices from the iterated best response* $\{B(k)\}$ $(k=1,2,3,\ldots)$ *consists of the maximal choice that survives BEWDSk under the MCM.*

Proof. See Appendix. ■

8.3. Coding Procedure

---

[24] This model is based on the level-k theory literature, while we apply iterated best response to the reduced normal form game. In the studies by Crawford and Iriberri (2007) and Costa-Gomes and Crawford (2006), subject in a static, dominance solvable game is said to be type $L1$ if he or she best responds to the type with the randomized choice, $L0$. Type $Lk+1$ is defined inductively as those who best respond to type Lk. Johnson et al. (2002) is a notable exception, which examined the level of reasoning in extensive form games, especially in six-stage bargaining games, but their classification was based on how many future rounds subjects truncate.

[25] The third, fourth, and fifth iterations result in choosing three, two, and one tokens, respectively.

In this subsection, we report the coding procedure in order to investigate which type of reasoning process subjects described in the questionnaire during and after the experiment, including diagonalization, regret minimization and iterated best response. This method is similar to those presented by Cooper and Kagel (2005), Brandts and Cooper (2007) and Chen and Chen (2011).[26]

The coding proceeded as follows. First, the authors separately read the responses both in the MAM and in the MCM. Then, they created 30 sentence-based categories from typical responses (Table 12). These categories include the responses seen only in the MAM, only in the MCM, and in both treatments and are broadly separated into three groups: behavioral description in the decision stage, behavioral description in the choice stage, and subjects' trend.

Two research assistants (hereafter coders) were independently instructed and each performed the coding for all 120 subjects in the MAM and MCM. The instructions for coders are included in Appendix E. The coders were required to read subjects' descriptions and determine, for each subject, to which categories his or her descriptions belong and to check all relevant categories (0=description does not fall into the category, 1=description falls into the category). We refer to a coder's binary decision as the *rating*. Coders never met face-to-face and no efforts were made to reconcile the differences in individual coding.

The above-described design is suitable for three reasons. First, setting as many as 30 categories including descriptions in both treatments allows us to analyze subjects' responses inclusively. Second, using fixed categories across treatments helps coders make unbiased decisions. Third, sentence-based categories can capture whether subjects describe the idea of diagonalization, regret minimization and iterated best response more precisely compared with the word-based categories typically used in the literature. As a check of validity, we also calculated the cross-coder correlation by category. The overall average cross-coder correlation was 0.450 and the average cross-coder correlation across the five most frequent categories was 0.421.[27]

Table 12 about here.

---

[26] Note that our experiments did not allow communication among subjects, while the above literature studied which specific words and attitudes in informal communication facilitated cooperation among group members.

[27] These values are comparable to the results presented by Cooper and Kagel (2005): the overall average cross-coder correlation in their paper was 0.388, while the average cross-coder correlation across the five most frequent categories was 0.570.

8.4. Coding Results for MAM sessions

For diagonalization, Categories 5, 10, and 14 were designed to capture the reasoning step (*D*-1) introduced in subsection 8.1.1, which can be expressed in several ways. Similarly, Category 8 refers to step (*D*-2). Category 12 was designed to capture regret minimization in the first stage.[28,29]

Result 7. *i) Categories 5, 10, and 14 for (D-1), Category 8 for (D-2), and Category 12 for regret minimization are the third, 12th, fourth, first, and 14th most frequent responses.*
ii) 96.7% of subjects are deemed to follow diagonalization by either coder. Among them, 38.3% of subjects are deemed to follow it by both coders.
iii) 16.7% of subjects are deemed to follow regret minimization by either coder. Among them, 3.3% of subjects are deemed to follow it by both coders.

Table 13 summarizes the coding results of the MAM for the alternative models mentioned above. The leftmost column specifies alternative models. The second column lists the category numbers that represent the idea of the alternative model. The next three columns present the distribution of subjects according to the total rating of both coders in the category and the last column shows the average ratings of both coders in the category along with the relative ranking in parentheses.[30] In Appendix F, Table A11 lists the average rating of all 30 categories.

Table 13 about here.

Let us begin with the categories for diagonalization. Category 8 for (D-2) ranks top among the 30 categories with an average rating of 0.767. Furthermore, Categories 5 and 14 for (D-1) rank third and fourth, respectively.[31] This result means that descriptions consistent with diagonalization are most frequently found in these responses. By contrast,

---

[28] We did not make the categories consistent with regret minimization in the decision stage because we could not find such responses from the questionnaires.

[29] When coders engaged in coding, Category 13 in Table 12 was classified as regret minimization. However, we found that the sentence of that category was too imprecise to capture regret minimization. Therefore, we do not consider Category 13 to describe regret minimization in the analysis.

[30] Given a treatment and category *k*, let n_i^k be the number of subjects rated as 1 by coder $i = 1, 2$. Since we have 60 subjects in one treatment, the average rating is $(n_1^k / 60 + n_2^k / 60) / 2$.

[31] The second highest average rating was Category 3, which says that the subject decides to approve or not to maximize his or her points.

regret minimization (Category 12) ranks 14th with an average rating of 0.100. Hence, regret minimization is deemed to be a relatively minor reasoning process among subjects.

We further count the number of subjects that received a rating of 1 from at least one coder in Categories 5, 10, and 14 (*D*-1) *and* from at least one coder in Category 8 (*D*-2). These subjects are deemed to be followers of diagonalization. This number reaches 58 out of 60 subjects (96.7%). Moreover, 23 out of 60 subjects (38.3%) received a rating of 1 by both coders in both (D-1) and (D-2) categories. By contrast, only 10 out of 60 subjects (16.7%) received a rating of 1 from either coder for regret minimization. Among them, both coders rated just two out of 60 subjects (3.3%) as followers of regret minimization. This result supports the notion that subjects considered diagonalization (albeit subconsciously) but did not consider regret minimization.

8.5. Coding Results for MCM Sessions

Categories 16 and 17 were designed to capture the first and second rounds of iterated best response in the first stage, respectively.

Result 8. *i) Categories 16 and 17 for iterated best response and Category 12 for regret minimization are the 12th, sixth, and 19th most frequent responses.*
ii) 21.7% of subjects are deemed to follow the first round of iterated best response by either coder. Among them, 6.7% of subjects are deemed to follow it by both coders. 45.0% of subjects are deemed to follow the second round of iterated best response by either coder. Among them, 23.3% of subjects are deemed to follow it by both coders.
iii) 11.7% of subjects are deemed to follow regret minimization by either coder. Among them, 3.3% of subjects are deemed to follow it by both coders.

Table 14 reports the coding results in the MCM. Because Category 12 ranks 19th for regret minimization with an average rating of 0.075, this explains only a proportion of subjects' choices and descriptions as well as it did in the MAM. Iterated best response also does not receive a majority.

Table 14 about here.

Among the 30 categories, Category 16 ranks 12th for the first-round iteration with an average rating of 0.142, while Category 17 comes ahead of Category 16 for the

second-round iteration (sixth with an average rating of 0.342). Nevertheless, it is worthwhile noting that Category 17 has the highest average rating among all categories in the first stage (see Table A11).

In the MCM, the highest and second highest average rankings are Categories 3 and 4, which concern the maximization of points in the second stage. The third highest average ranking is Category 28, which concerns the adjustment of the contribution given the previous choice of the counterpart.

By counting subjects that received a rating of 1 from at least one coder and from both coders, 11.7% (=(5+2)/60) and 3.3% (=2/60) of subjects fall into Category 12 for regret minimization, respectively. Moreover, 21.7% (=(9+4)/60) and 6.7% (=4/60) of subjects fall into Category 16 for the first round of iterated best response, while 45.0% (=(13+14)/60) and 23.3% (=14/60) of subjects fall into Category 17 for the second round of iterated best response, respectively.

8.6. Inequality Aversion in the Second Stage of the MCM

As shown in subsection 7.3.2, we observed weakly dominated disapproval in the second stage of the MCM. We conjecture that this result is brought about by inequality aversion. We thus recalculate the coding results when a subject with a higher first-stage choice than the other player chose n even though the BEWDS prediction is y. We find that 60.6% of decisions in this situation are given by either coder a positive rating in Category 1 or 7, which is consistent with the notion of inequality aversion.[32] Further, 28.2% of decisions are undetermined because neither coder rated subjects. It is remarkable that 34 out of the 71 weakly dominated disapprovals introduced in subsection 7.3.2 occur because of the groups that included the four subjects in the first session who kept choosing 24 tokens and disapproved the other subject's choice in most periods. When we focus on these four subjects, we obtain a similar result, namely that 60% of decisions are given a rating of 1 by either coder and that the remaining decisions are undetermined. This result lends support to the argument presented in subsection 7.3.2.

9. Concluding Remarks

This paper theoretically and experimentally showed that the MAM leads to the efficient provision of public goods within a standard linear public goods game framework.

[32] Coders recorded the period or questionnaire in which the subject's description is relevant to the category. If the decision is not rated as Category 1 or 7 by a coder, but the description is the same as that given a rating of 1 in earlier periods, we counted it as having received a rating of 1. If the decision is given a rating of 1 in Category 1 or 7, but is also given a rating of 1 in other categories in the decision stage that are inconsistent with inequality aversion, we did not count it as having received a rating of 1.

One important theoretical departure from the literature is implementation in BEWDS rather than in dominant strategy equilibria, NE, or SPNE. In addition, the MAM has a distinctive feature of voluntariness compared with the mechanisms typically proposed in the literature that rely on coercive tax power (see Groves and Ledyard, 1977; Falkinger, 1996).

In the presented experiments, we observed rapid convergence towards the efficient allocation in MAM sessions. Although our environmental setting is simple, our result is in stark contrast to recent studies of public good mechanisms, which have often needed dozens of repetitions to attain efficiency in their experiments (see Chen and Plott, 1996; Andreoni and Varian, 1999; Chen and Gazzale, 2004; Healy, 2006). By using two criteria in order to compare the predictive performances of equilibrium concepts in both stages and in the second stage only, we found that subjects consistently chose BEWDS. Indeed, SPNE did not significantly improve the explanation of our data relative to BEWDS in all treatments. Further, our inclusive coding of subjects' responses helped to pinpoint their reasoning processes which produce outcomes consistent with BEWDS. Beyond the authors' expectations, the majority of subjects in the MAM explicitly described the idea of diagonalization. We also confirmed that certain subjects followed regret minimization both in the MAM and in the MCM, whereas they followed iterated best response and inequality aversion in the MCM.

One direction for further research is to evaluate experimentally BEWDS implementation mechanisms including approval mechanisms in more general environments. As noted in Section 3, when there are only two players that have different endowments and preferences to the public good, we can almost directly apply the MAM by reducing the problem to the symmetric case through asking players to choose the ratio of contribution to his or her endowment. Using homogeneous n-person public good environments with binary choices, Huang et al. (2012) succeeded in constructing a variant of an n-person approval mechanism. However, the SMAM introduced in this paper extends the mechanism of Huang et al. (2012) to homogeneous n-person public good environments with continuous contributions and linear payoffs by selecting players who proceed to the second stage. Future work might also compare BEWDS implementation mechanisms with existing Nash or subgame perfect Nash implementation mechanisms in order to shed new light on the rationality assumptions that underlie the design of public good mechanisms.

This paper should at least stimulate constructive discussion on the theory of mechanism designs that allow players to have heterogeneous reasoning processes including BEWDS. The observed heterogeneity is in line with previous experimental

research not only in terms of social preferences (see Cooper and Kagel, 2008 for a survey) but also in terms of strategic sophistication (Crawford and Iriberri, 2007) and learning with forgone payoffs (Ho et al., 1998). Even though our experiment used a simple public good environment, this paper has important implications for practical mechanism design, which should emphasize the alignment of players with various behavioral rules towards a social optimum.

Acknowledgements

We would like to thank Koji Abe, Kanemi Ban, Vincent Crawford, Yukihiko Funaki, Michihiro Kandori, Tetsuya Kawamura, Yukio Koriyama, Shuhei Morimoto, Yuko Morimoto, Alistair Munro, Yasutomo Murasawa, Masuyuki Nishijima, Masahiro Okuno, Dimitry Rtischev, Yoshinao Sahashi, Tatsuhiro Shichijo, Mitsuru Sunada, Masanori Takaoka, Yoshiyuki Takeuchi, Masanori Takezawa, Wataru Tamura, Hisashi Tanizaki, Hiroshi Uno, Toshio Yamagishi, and Takafumi Yamakawa. We also thank the participants at the 11th SAET Conference, the 2011 Japanese Economic Association Autumn Meeting, the 2012 Asia-Pacific Economic Science Association Meeting, and the seminar at Osaka Prefecture University for their helpful comments. This research was supported by the Joint Usage/Research Center at ISER, Osaka University and "Experimental Social Sciences: Toward Experimentally-based New Social Sciences for the 21st Century," which is a project called the Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, and Culture of Japan. The first author acknowledges the financial support from the Research Fellowships for Young Scientists of the Japan Society for the Promotion of Science.

References

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

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

Attiyeh, G., Robert, F., Isaac, R.M., 2000. Experiments with the pivot process for providing public goods. Public Choice 102 (1), 93-112.

Banks, J.S., Plott, C.R., Porter, D.P., 1988. An experimental analysis of unanimity in public

goods provision mechanisms. Rev. Econ. Stud. 55 (2), 301-322.

Banks, J.S., Camerer, C., Porter, D.P., 1994. An experimental analysis of Nash refinements in signaling games. Games. Econ. Behav. 6 (1), 1-31.

Binmore, K., McCarthy, J., Ponti, G., Samuelson, L., Shaked, A., 2002. A backward induction experiment. J. Econ. Theory 104 (1), 48-88.

Brandts, J., Holt, C.A., 1992. An experimental test of equilibrium dominance in signaling games. Am. Econ. Rev. 82 (5), 1350-1365.

Brandts, J., Cooper, D.J., 2007. It's what you say, not what you pay: an experimental study of manager-employee relationships in overcoming coordination failure. J. Eur. Econ. Assoc. 5 (6), 1223–1268.

Cason, T.N., Saijo, T., Sjostrom, T., Yamato, T., 2006. Secure implementation experiments: do strategy-proof mechanisms really work? Games Econ. Behav. 57 (2), 206–235.

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

Chen, R., Chen, Y., 2011. The potential of social identity or equilibrium selection. Amer. Econ. Rev. 101 (6), 2562-2589.

Chen, Y., Gazzale, R., 2004. When does learning in games generate convergence to Nash equilibrium? The role of supermodularity in an experimental setting. Amer. Econ. Rev. 94 (5), 1505–1535.

Chen, Y., Plott, C.R., 1996. The Groves-Ledyard mechanism: an experimental study of institutional design. J. Public Econ. 59, 335-364.

Chen, Y., Tang, F., 1998. Learning and incentive compatible mechanisms for public goods provision: an experimental study. J. Polit. Economy 106 (3), 633–662.

Cooper, D., Kagel, J., 2009. Other regarding preferences: a selective survey of experimental

results, in: Kagel J., Roth, A. (Eds.), The Handbook of Experimental Economics, volume 2., In Press.

Cooper, D., Kagel, J., 2008. Learning and transfer in signaling games. Econ. Theory 34(3), 415-439.

Cooper, D.J., Kagel, J.H., 2005. Are two heads better than one? Team versus individual play in signaling games. Amer. Econ. Rev. 95, 477–509.

Costa-Gomes, M.A., Crawford, V.P., 2006. Cognition and behavior in two-person guessing games: An experimental study. Amer. Econ. Rev. 96 (5), 1737-1768.

Crawford, V.P., Iriberri, N., 2007. Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? Econometrica 75 (6), 1721-1770.

Croson, R.T.A., 2008. Public goods experiments, in: Durlauf, S.N., Blume, L.E. (Eds.), The New Palgrave Dictionary of Economics, Second Ed. Palgrave Macmillan, London.

Erev, I., Roth, A., 1998, Predicting how people play games: reinforcement learning in games with unique strategy equilibrium. Amer. Econ. Rev. 88 (4), 848-881.

Ert, E., Erev, I., Roth, A.E., 2011. A choice prediction competition for social preferences in simple distribution games: an introduction. Games 2 (3), 257-276.

Falkinger, J., 1996. Efficient private provision of public goods by rewarding deviations from average. J. Public Econ. 62 (3), 413-422.

Falkinger, J., Fehr, E., Gachter, S., Winter-Ebmer, R., 2000. A simple mechanism for the efficient provision of public goods: experimental evidence. Amer. Econ. Rev. 90 (1), 247-264.

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

Green, J., Laffont, J-J., 1977. Characterization of satisfactory mechanisms for the revelation

of preferences for public goods. Econometrica 45 (2), 427-38.

Groves, T., Ledyard, J., 1977. Optimal allocation of public goods: a solution to the "free rider" problem. Econometrica 45 (4), 783-809.

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

Huang, X., Masuda, T., Okano, Y., Saijo, T., 2012. Toward solving social dilemma: theory and experiment. In preparation.

Healy, P.J., Mathevet, L., 2012. Designing stable mechanisms for economic environments. Theoretical Econ. 7, 609–661.

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

Ho, T.H., Camerer, C., Weigelt, K., 1998. Iterated dominance and iterated best response in experimental "p-beauty contests". Amer. Econ. Rev. 88 (4), 947-969.

Johnson, E.J., Camerer, C., Sen, S., Rymon, T., 2002. Detecting failures of backward induction: monitoring information search in sequential bargaining. J. Econ. Theory 104 (1) 16-47.

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

Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popovic, Z., Jaskolski, M., Baker, D., 2011. Crystal structure of monomeric retroviral protease solved by protein folding game players. Nat. Struct. Mol. Biol. 18, 1175–1177.

Laffont, J-J., Maskin, E., 1982. Nash and dominant strategy implementation in economic environments. J. Math. Econ. 10 (1), 17-47.

Ledyard, J.O., 1995. Public goods: a survey of experimental research, in: Kagel J., Roth, A.

(Eds.), The Handbook of Experimental Economics, Princeton University Press, Princeton, pp. 111–194.

Nyarko, Y., Schotter, A., 2002. An experimental study of belief learning using elicited beliefs. Econometrica 70 (3), 971-1005.

Renou, L., Schlag, K.H., 2011. Implementation in minimax regret equilibrium. Games Econ. Behav. 71 (2), 527-533.

Saijo, T., Yamato, T., Sjöström, T., 2007. Secure implementation. Theoretical Econ. 2 (3), 203-229.

Saijo, T., Okano, Y., Yamakawa, T., 2012. The mate choice mechanism experiment: a solution to prisoner's dilemma. Unpublished manuscript.

Samuelson, P.A., 1954. The pure theory of public expenditure. Rev. Econ. Statist. 36 (4), 387–389.

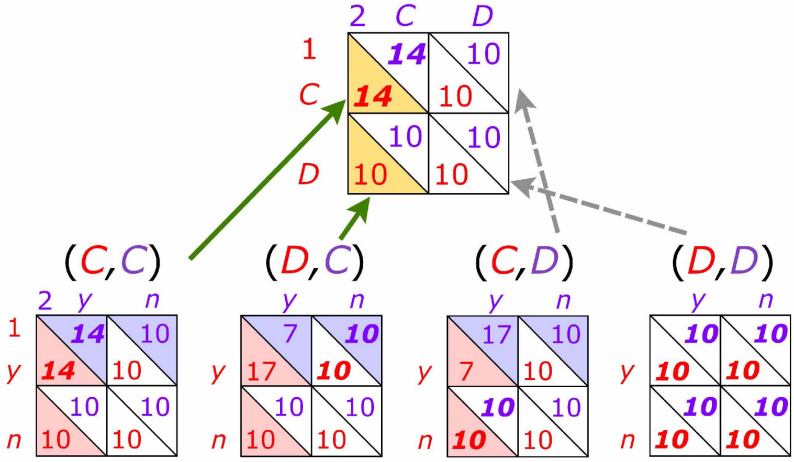
Smith, V.L., 1979. An experimental comparison of three public good decision mechanisms. Scand. J. Econ. 81 (2), 198-215.

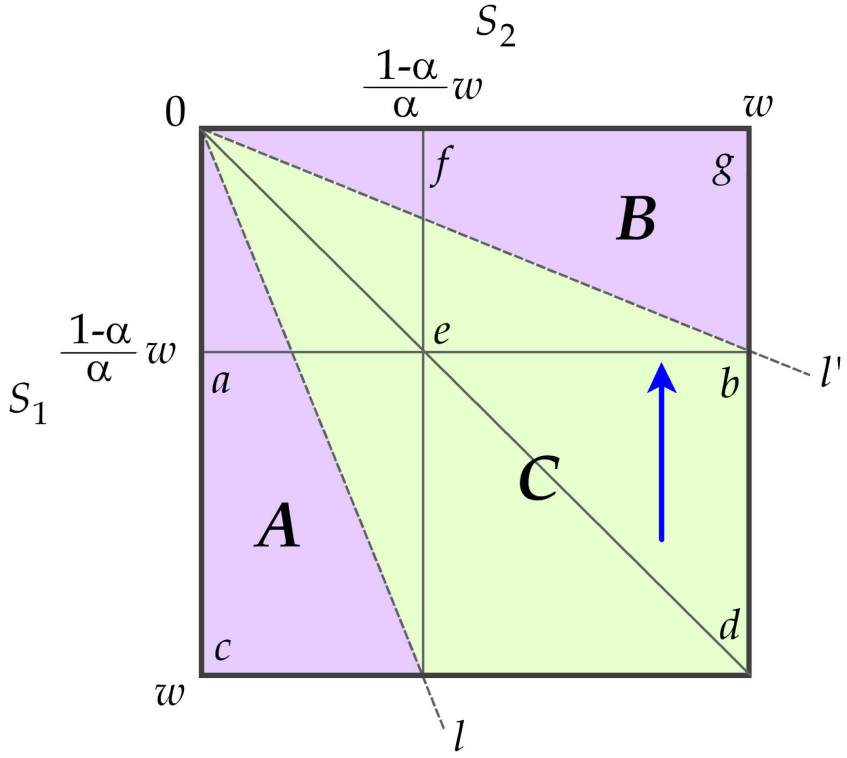
Smith, V.L., 1980. Experiments with a decentralized mechanism for public good decisions. Amer. Econ. Rev. 70 (4), 584-599.

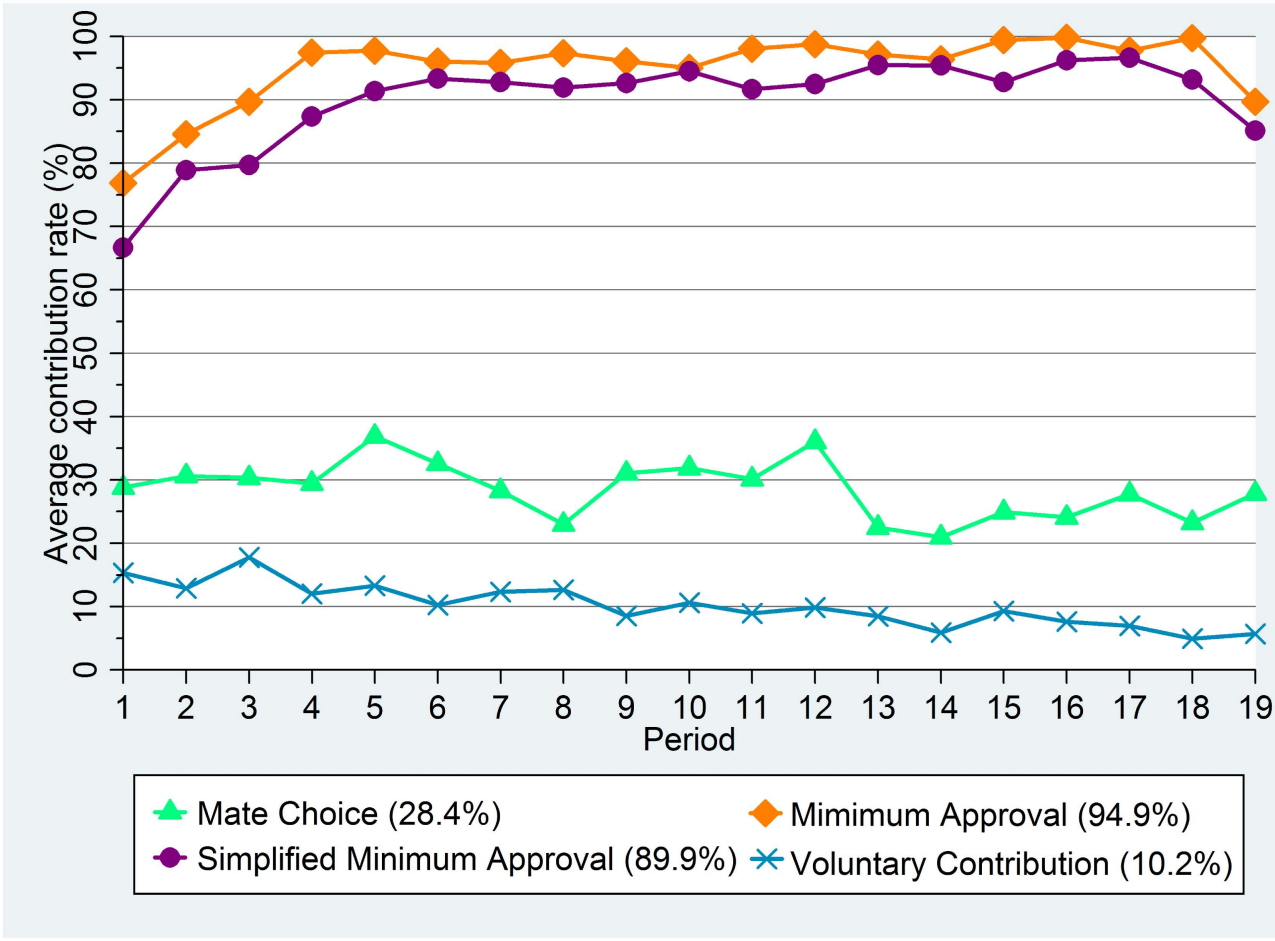
Stahl, D., Haruvy, E., 2009. Testing theories of behavior for extensive-form two-player two-stage games. Exper. Econ. 12 (2), 242-251.

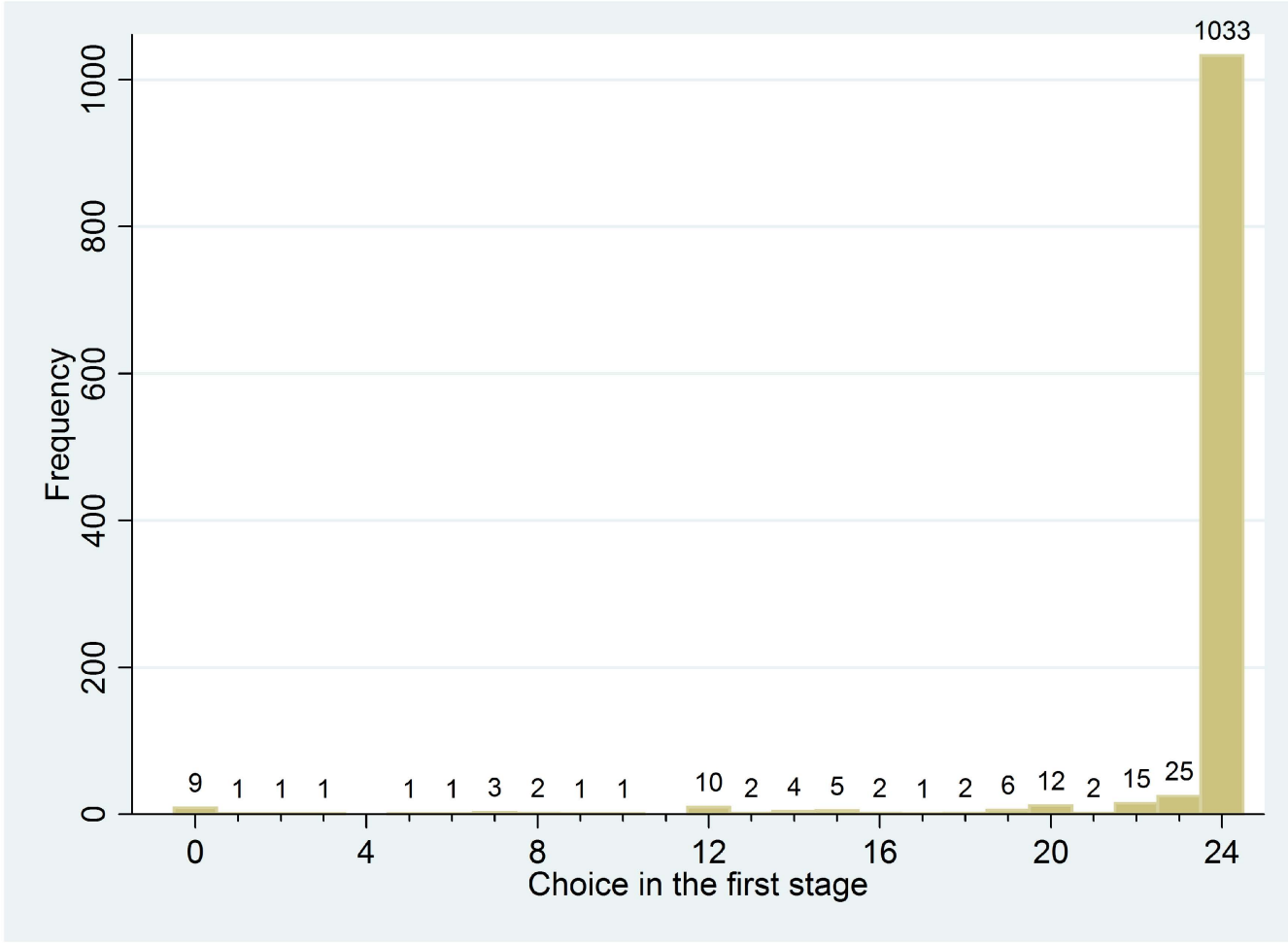
Varian, H., 1994. A solution to the problem of externalities when agents are well-informed. Amer. Econ. Rev. 84 (5), 1278-1293.

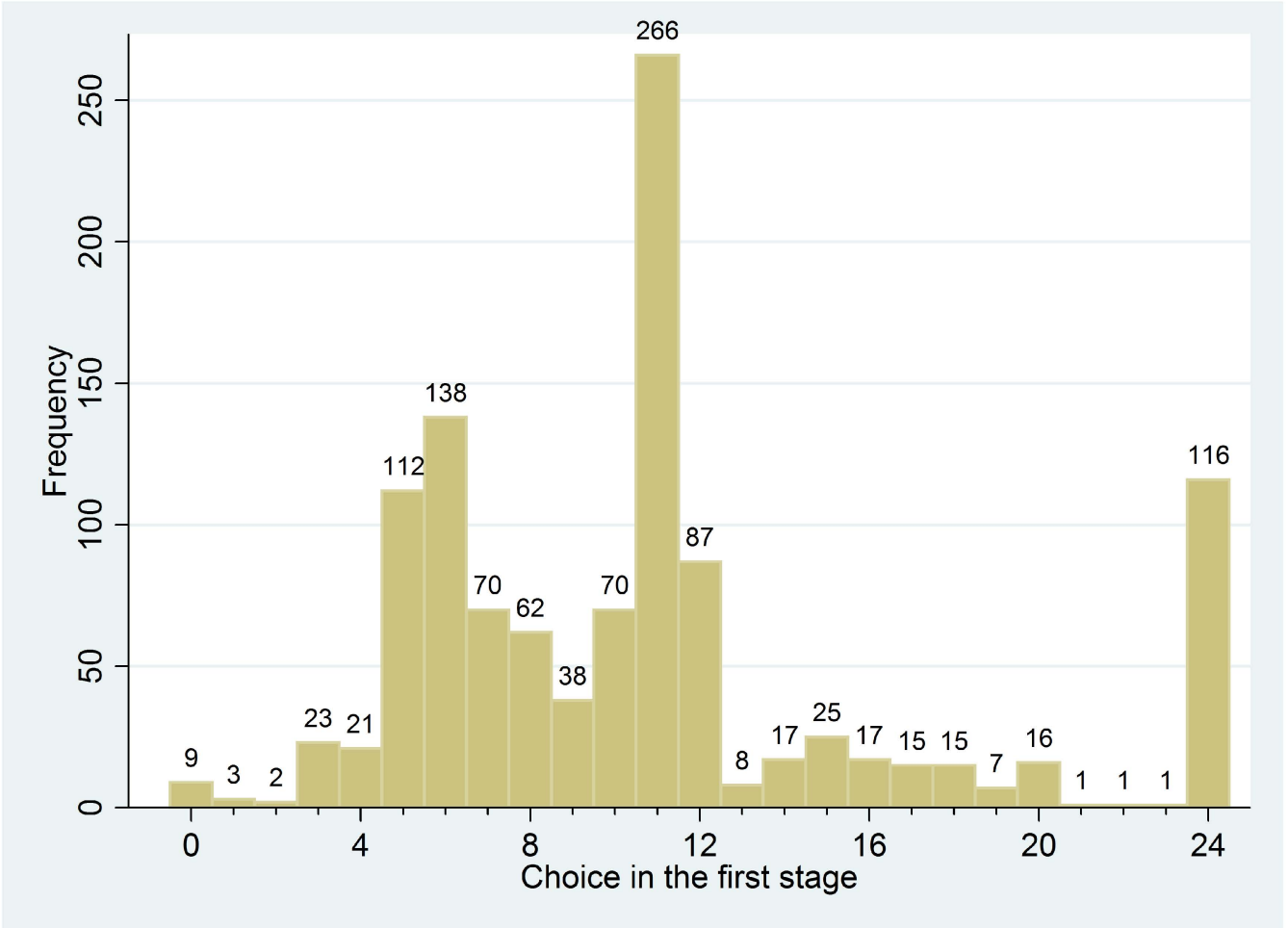
Walker M., 1981. A simple incentive compatible scheme for attaining Lindahl allocations. Econometrica 49, 65–71.


2 C D
1
C
D
14
14
10
10
10
10
10
10
(C,C)
(D,C)
(C,D)
(D,D)
2 y n
1
y
n
14
14
10
10
10
10
10
10
y n
y
n
7
17
10
10
10
10
y n
y
n
17
7
10
10
10
10
y n
y
n
10
10
10
10
10
10


S_2
$\frac{1-\alpha}{\alpha}w$
0
w
f
g
B
$\frac{1-\alpha}{\alpha}w$
S_1
e
l'
a
b
C
A
d
c
w
l


Average contribution rate (%)
0
10
20
30
40
50
60
70
80
90
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Period
Mate Choice (28.4%)
Mimimum Approval (94.9%)
Simplified Minimum Approval (89.9%)
Voluntary Contribution (10.2%)


Frequency
0
200
400
600
800
1000
9
1
1
1
1
1
3
2
1
1
10
2
4
5
2
1
2
6
12
2
15
25
1033
0
4
8
12
16
20
24
Choice in the first stage


Frequency
0
50
100
150
200
250
9
3
2
23
21
112
138
70
62
38
70
266
87
8
17
25
17
15
15
7
16
1
1
1
116
0
4
8
12
16
20
24
Choice in the first stage

	0	2	4	6	8	10
0	10	11.4	12.8	14.2	15.6	17
2	9.4	10.8	12.2	13.6	15	16.4
4	8.8	10.2	11.6	13	14.4	15.8
6	8.2	9.6	11	12.4	13.8	15.2
8	7.6	9	10.4	11.8	13.2	14.6
10	7	8.4	9.8	11.2	12.6	14

Table 1. Player 1's payoff table when w=10 and $\alpha = 0.7$.

	y	*n*
y	(14.2,8.2)	(10,10)
n	(10,10)	(10,10)

Subgame (0,6)

	y	*n*
y	(15.2,11.2)	(10,10)
n	(10,10)	(10,10)

Subgame (6,10)

Table 2. Payoff tables in subgames (0,6) and (6,10).

	0	2	4	6	8	10
0	10	10	10	10	10	10
2	10	10.8	12.2	10	10	10
4	10	10.2	11.6	13	14.4	10
6	10	10	11	12.4	13.8	15.2
8	10	10	10.4	11.8	13.2	14.6
10	10	10	10	11.2	12.6	14

Table 3. Player 1's payoff table in the reduced normal form game under the MCM.

	y	*n*
y	(9.6,13.6)	(10.8,10.8)
n	(10.8,10.8)	(10.8,10.8)

Subgame (6,2)

	y	*n*
y	(14.6,12.6)	(13.2,13.2)
n	(13.2,13.2)	(13.2,13.2)

Subgame (8,10)

Table 4. Payoff tables in subgames (6,2) and (8,10).

	0	2	4	6	8	10
0	10	10	10	10	10	10
2	10	10.8	10.8	10.8	10.8	10.8
4	10	10.8	11.6	11.6	11.6	11.6
6	10	10.8	11.6	12.4	12.4	12.4
8	10	10.8	11.6	12.4	13.2	13.2
10	10	10.8	11.6	12.4	13.2	14

Table 5. Player 1's payoff table in the reduced normal form game under the MAM.

	0	2	4	6	8	10
0	10	10.4	10.8	11.2	11.6	12
2	10.4	10.8	11.2	11.6	12	12.4
4	10.8	11.2	11.6	12	12.4	12.8
6	11.2	11.6	12	12.4	12.8	13.2
8	11.6	12	12.4	12.8	13.2	13.6
10	12	12.4	12.8	13.2	13.6	14

Table 6. Player 1's payoff table in the reduced normal form game under the AAM.

	y	*n*
y	(13.8,11.8)	(10,10)
n	(10,10)	(10,10)

Subgame (8,10)

	y	*n*
y	(14.4,10.4)	(10,10)
n	(10,10)	(10,10)

Subgame (4,8)

Table 7. Payoff tables in subgames (8,10) and (4,8).

Table 8. Two-tailed Mann–Whitney test statistics for the equality of average contributions, grouped by treatment.

Treatment	*SMAM*	*MCM*	*VCM*
MAM	3.799**	9.450**	9.445**
SMAM		9.133**	9.275**
MCM			7.284**

Notes: **p<0.01.

(%)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
MAM	-	-	-	-	-	-	-	-	13	47
SMAM	1	-	-	-	-	1	-	2	15	41

Table 9. Distribution of average contribution rates per subject

Treatment	Equil. concept	Equil. paths	# of equil. paths	# of equil.-consistent path data	McNemar's exact p-value
MAM	BEWDS	$((24, \cdot),(24, \cdot))$[a]	4	475	0.250
	SPNE	$((s \cdot)(s,\cdot))$, s =0,1,…,24	100	478	
SMAM	BEWDS	(24, 24)	1	424	1.000
	SPNE	(s,s), s =0,1,…,24	25	425	
MCM	BEWDS	$((s_1,t_1),(s_2,t_2))$, s_1,s_2=1,2,…,11[b),c)]	123	264	<0.000
	SPNE	$((s_1,v_1),(s_2,v_2))$, s_1,s_2=0,1,…,24[d)]	1160	495	

Notes : a) Since subjects in the MAM proceed to the decision stage when both players choose the same choice in the first stage, there are four payoff equivalent BEWDS paths. On the other hand, subjects in the SMAM does not proceed to the decision stage in such cases and hence there is a unique BEWDS path. b) This prediction is based on BEWDS1.

c)

$$(t_1,t_2)=\begin{cases}(n,y)\text{ if } s_1>0.7s_2/0.3\\ (y,y)\text{ or }(n,y)\text{ if }(s_1,s_2)=(7,3),(14,6),(21,9)\\ (y,y)\text{ if } s_1<0.7s_2/0.3 \text{ and } s_2<0.7s_1/0.3\\ (y,y)\text{ or }(y,n)\text{ if }(s_1,s_2)=(3,7),(6,14),(9,21)\\ (y,n)\text{ if } s_2>0.7s_1/0.3\\ (y,y),\ (y,n),\ (n,y)\text{ or }(n,n)\text{ if }(s_1,s_2)=(0,0).\end{cases}$$

d)

$$(v_1,v_2)=(t_1,t_2)\text{ or }(n,n)\text{ for all }(s_1,s_2).$$

Table 10. Evaluation of BEWDS and SPNE by treatment.

Treatment	Subgames[a), b)]	The second-stage decisions				# of equil. decisions		% of subgame-consistent path data		McNemar's exact p-value
						BEWDS/ SPNE		BEWDS: shaded cells	SPNE: shaded cells+(n,n)	
MAM		$(y, \boldsymbol{y})$	$(n, \boldsymbol{y})$	(y,n)	(n,n)					
	$s_1 > s_2$	9	79	0	4	1	2	85.9	90.2	0.125
SMAM		$(y,-)$[c)]	$(n,-)$							
	$s_1 > s_2$	4	141			1	2	97.2	97.2	1.000
MCM		$(y, \boldsymbol{y})$	$(n, \boldsymbol{y})$	(y,n)	(n,n)					
	$s_1 < 0.7s_2/0.3$	377	71	0	1	1	2			
	$s_1 = 0.7s_2/0.3$	2	2	0	0	2	3	86.3	86.8	0.250
	$s_1 > 0.7s_2/0.3$	4	111	0	2	1	2			

Notes: a) Without loss of generality we can assume $s_1 \geq s_2$. b) The path data where both subjects are indifferent between approval and disapproval are omitted. c) "–" indicates that player 2 does not proceed to the decision stage in the SMAM. d) n is always the unique best response in the SMAM. e) Bold y indicates that player 2 chooses y under BEWDS.

Table 11. Subgame-consistent path data under BEWDS and SPNE by treatment.

Large category	Category No.	Category and discriptions
Decision stage	1	(One's own motives 1) I disapprove when I lose by a wide margin.
	2	(The counterparts' motives 1) Others disapprove when they lose by a wide margin.
	3	(One's own motives 2) I decide to approve or not to maximize my points.
	4	(The counterparts' motives 2) Others decide to approve or not to maximize their own points.
	5	(Linking to approval or not with first-stage choices) Whenever two players choose different investment in the choice stage, either will disapprove to maximize his or her own points.
	6	(Trying to make a good impression) When approval and disapprovals are indifferent, I approve so that other subjects have a good impression of me.
	7	(Asking for cooperation even if points are lost) I choose 24 tokens (or slightly smaller number) to maximize the sum of points of the pair. If my counterpart chooses an even smaller number, I disapprove even if I lose my points.
Choice stage	8	(Profit maximization under symmetric contributions) If both players contribute the same amount, I can maximize my points by choosing 24 tokens.
	9	(Avoiding being disapproved) I choose investments to be approved, no matter how much points I get.
	10	(Referring to the diagonal line) I can get points only on the diagonal line of the points table.
	11	(Expected utility maximization) I choose investments to maximize my expected points, considering the past observations of my counterpart's strategies.
	12	(Regret minimization 1) I made risk-free choices to lose points no matter what others choose in the choice stage.
	13	(Regret minimization 2) I made the first-stage choice to be approved no matter what others choose in the choice stage.
	14	(Difference in points earned between two players) Both end up receiving the same points.
	15	(Optimistic prediction) Even when I slightly decrease my investment, my counterpart will approve.
	16	(Taking advantage of a cooperative subject 1) My counterpart will choose approximately 24 tokens. Then, I can maximize my points by choosing 11 tokens (or a slightly larger number) in the choice stage given long as my counterpart approves.
	17	(Taking advantage of a cooperative subject 2) My counterpart will choose approximately 11 tokens. Then, I can maximize my points by choosing five tokens (or a slightly larger number) in the choice stage as long as my counterpart approves.
	18	(Eliminating weakly dominated strategies) No matter what investment others choose, I would be better off by choosing 11 tokens compared with choosing larger numbers.
	19	(Trying to affect other subjects) Others should increase investments as I do.
Trend	20	(Expectation 1) I always anticipated what my counterpart would do in the choice stage.
	21	(Expectation 2) I always anticipated what my counterpart would do in the decision stage.
	22	(Expectation 3) Contributions tend to decrease.
	23	(Expectation 4) Contributions will decrease until one.
	24	(Imitation) I mimicked others' behavior.
	25	(Dilemma) While I know that choosing large numbers is mutually beneficial, we cannot achieve it.
	26	(Estimation from past observations) Since counterparts adopt various strategies, it is important to estimate the distribution of their strategies.
	27	(Reaction of the counterpart in the last period 1) Since I was disapproved in the last period, I increased my investment in this period.
	28	(Reaction of the counterpart in the last period 2) Since I was approved in the last period, I decreased my investment in this period.
	29	(Subjects who think differently) It is surprising that some subjects disapprove at the expense of their own points.
	30	(Subjects who think similarly) I confirmed that other subjects think the same way as I do.

Table 12. The full list of coding categories.

Alternative model		Category number	Number of subjects with total rating of			Average rating (ranking)
			0	1	2	
Diagonalization	Step D-1	5	4	30	26	0.683 (3)
		10	48	7	5	0.142 (12)
		14	19	29	12	0.442 (4)
	Step D-2	8	0	28	32	0.767 (1)
Regret minimization		12	50	8	2	0.100 (14)

Table 13. Coding results for alternative models for the MAM.

Alternative model	Category number	Number of subjects with total rating of			Average rating (ranking)
		0	1	2	
Regret minimization	12	53	5	2	0.075 (19)
Iterated best response	16	47	9	4	0.142 (12)
	17	33	13	14	0.342 (6)

Table 14. Coding results for alternative models for the MCM.