WINPEC Working Paper Series No.E1604
June 2016

Solving the second-order free rider problem in a public goods game: An experiment using a leader support system

Hiroki Ozono, Nobuhito Jin, Motoki Watabe and Kazumi Shimizu

Waseda INstitute of Political EConomy
Waseda University
Tokyo,Japan

Title: Solving the second-order free rider problem in a public goods game: An experiment using a leader support system

Hiroki Ozono[a,*], Nobuhito Jin[b], Motoki Watabe[c], and Kazumi Shimizu[d]

[a]Faculty of Law, Economics and Humanities, Kagoshima University, 1-21-30, Korimoto, Kagoshima 890-0065, JAPAN

[b]School of Psychology Practices, College of Integrated Human and Social Welfare Studies, Shukutoku University, 200, Daiganji-cho, Chuo-ku, Chiba-Shi, 260-8701, JAPAN

[c] School of Business, MonashUniversity, Malaysia, Jalan Lagoon Selatan, 46150 Bandar Sunway, Selangor Darul Ehsan, MALAYSIA

[d]School of Political Science and Economics, Waseda University, Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, JAPAN

[*]Corresponding Author:

Hiroki Ozono

Faculty of Law, Economics and Humanities, Kagoshima University,

Korimoto, Kagoshima 890-0065, JAPAN

+81-99-285-7538

Hiroki.ozono@gmail.com

Abstract:

To study the collective action problem, researchers have investigated public goods games (PGG), in which each member decides to contribute to a common pool that returns benefits to all members equally. Punishment of non-cooperators—free riders—can lead to high cooperation in PGG. However, the existence of second-order free riders, who do not pay punishment costs, reduces the effectiveness of punishment. We focus on a "leader support system," in which one group leader can freely punish group followers using capital pooled through the support of group followers. In our experiment, participants were asked to engage in three stages: a PGG stage in which followers decided to cooperate for their group; a support stage in which followers decided whether to support the leader or not; and a punishment stage in which the leader could punish any follower. We found both higher cooperation and higher support for a leader achieved under linkage-type leaders—who punished both non-cooperators and non-supporters. In addition, linkage-type leaders themselves earned higher profits

than other leader types because they withdrew more support. This means that a leader who effectively punishes followers could increase their own benefits and the second-order free rider problem would be solved.

Key words: Cooperation, Punishment, Leadership, Public goods, Governance

1. Introduction

1.1. Public goods problem and peer punishment

The difficulties of constructing a cooperative relationship have been formulized as a public goods problem [1, 2] and many studies have been conducted in the social sciences. In a typical public goods game (PGG), group members decide how much of their own resources to contribute to the common pool and the resources gathered in the pool benefit members equally. The entire group earns the highest profit when all members contribute all of their resources. Free riders, however, can increase their own payoff if they contribute nothing and still benefit from the common pool. This results in a socially inefficient situation. This is called the free rider problem. Humans encounter many PGG situations on a daily basis. In historical times, there were PGG situations involving food distribution in a hunter–gatherer society and irrigation facility work in an agrarian society. In modern times, such PGG situations range widely from

small-scale issues, like housework distribution within a household, to large-scale ones, such as an act to prevent a global-warming effect.

Peer punishment is one of the solutions that have been proposed by scholars [3–6]. This involves individuals punishing free riders, decreases the incentive to free ride and, thus, establishes cooperation [7]. Despite several laboratory experiments indicating that peer punishment solves the free rider problem [3–6], several theoretical and empirical questions have been posed. The biggest theoretical issue is the second-order free rider problem [8, 9]. Because punishing someone incurs cost, owing the punishment cost to maintain group cooperation is a second-order cooperative action. However, because the individuals' profits increase if s/he punishes nobody, second-order free riders emerge. Thus, theoretically speaking, peer punishment should not evolve. The reasons why people punish even if they have to pay the cost have been proposed using multilevel selection theory [10] and reputational benefits for punishers [11.12]; however, both theories have received criticism ([13] for multi-level selection; [14] for reputation for punishers) and there is no sufficient answer yet. Furthermore, an anthropological survey showed that punishment between individuals is rare in a small society, which is similar to an evolved environment [15]. As such, there are several doubts as to whether peer punishment solely can solve the public goods problem.

1.2. Pool punishment system as a solution to the public goods problem

Other than the peer punishment system, the pool punishment system has been proposed to solve the public goods problem [16–20]. Sigmund et al. (2010) [18] compared peer punishment with pool punishment, where group members pay cost to a punishment-executing system (e.g., a police force) and the system uses these resources as capital to punish the free riders. The authors mathematically showed that the pool punishment system is more stable than peer punishment only when the system punishes not just the first-order free rider but also the second-order free riders, who do not bear the cost of the punishment system. Traulsen et al. (2012) [20] examined the pool punishment system by implementing it in a laboratory experiment. Their experiment showed that participants tended to select pool punishment rather than peer punishment. In addition, they reported that systems with second-order punishment increased the number of people bearing the punishment cost, and so, high cooperation was likely to be achieved compared to the condition with only first-order punishment.

In our study, we postulated the executor of pool punishment as a leader. In previous studies on pool punishment, the executor of punishment is assumed to be a system governed by all group members and punishment is executed automatically in accordance with its rules. Thus,, the profit of the punishment system itself was not considered. Indeed, these systems have existed in actual societies [21]. In many cases,

however, each pool system was governed by a leader or a few leaders, such as headmen in villages, lords of manors, or kings in nations. In such systems, followers support their leaders by giving their own resources to their leaders. We name these the "leader support system," a leader can freely decide the punishment rules: whom to punish and to what degree. In addition, it should be considered that the leader can obtain the surplus resources that were pooled by the followers but were not used as the cost of punishment. In previous studies on pool punishment, no one can obtain surplus resources because they should be used to maintain the pool punishment system even if all members were cooperators [18–20]. Although maintenance cost is necessary, we consider that the surplus resources should remain if enough resources are pooled to the leader. These surplus resources bring the leader the incentive to induce support from followers. Under the leader support system, a leader can obtain more support and profit by executing not only the first-order but also the second-order punishment, which also results in high cooperation in PGG. Considering the group leader as a punishment executor makes us possible to reveal the origin of the punishment rule; how and why did the second-order punishment emerge?

1.3. Who does a leader punish in the leader support system?

Our study aimed to address the issues within the pool punishment system by setting up

a leader support system. Under this system, a specific person, namely a group leader, can freely punish group followers using resources pooled through support from group followers as capital. Specifically, in a group experiment conducted in a laboratory, one participant is assigned as the leader and the other participants as group followers. The participants were asked to engage in three stages: a PGG stage in which all the followers engage in PGG; a support stage in which the followers have opportunity to provide support for the leader; and a punishment stage in which the leader can freely punish followers. Our study examined the behavior of the leader and followers and the group cooperation level in a PGG.

When a leader support system is postulated, first, an incentive must be provided for the leader to punish those who do not support him/her (termed "self-focused punishment"). This is because the long-term profit of the leader him/herself may increase by changing the behavior of non-supporters into supporters by punishing them. Under such a self-focused punishment leader, non-cooperators in PGG are not punished, and so, high cooperation will not be attained and a tyrannical state might arise in which the followers continue to support the leader out of fear of punishment. In other words, there might arise a distorted state in which second-order free riders are punished but first-order free riders are not under a self-focused punishment leader.

Second, an incentive should be generated for the leader to punish non-cooperators in the PGG (termed "group-focused punishment") in order to gain the trust of the followers. This is because the leader's profit is expected to increase in the long term if s/he is recognized as an effective leader by the followers and draws out the support of followers. Although there is a possibility that high PGG cooperation can be achieved under a group-focused punishment leader, it would be difficult to garner support because non-supporters are not punished. Thus, there is a risk that resources for punishment may eventually be insufficient. This means that second-order free riders will increase because a group-focused punishment leader punishes only first-order free riders. Thus, leaders are not able to punish first-order free riders easily, making it difficult to maintain high PGG cooperation. As shown here, achieving high cooperation in the PGG under both self-focused and group-focused punishment was predicted to be difficult.

Then, what happens if a "linkage punishment" leader—who executes both self-focused and group-focused punishment—appears? Because the second-order free riders are punished, such leaders can gather support. A high cooperation in the PGG can be achieved easily considering that the leader can punish the first-order free riders using gathered support as capital. In addition, a linkage punishment leader would be

more likely to gain support from the followers compared to self-focused punishment leaders because s/he is beneficial to the group and gains followers' trust. Thus, the profit of the leader him/herself is expected to increase as well. The core problem of the second-order free rider is that there is no incentive to pay a punishment cost. For the followers, however, the incentive to bear the punishment cost (i.e., to support the leader) is provided to evade punishment from the leader. For the leader, the incentive to bear the first-order and second-order punishment costs (i.e., punishment toward non-cooperative and non-supporters, respectively) is generated to attain support from followers in the long term. Therefore, the second-order free rider problem can be solved under the leader support system if a linkage punishment leader appears. This subsequently creates a state in which both the followers and the leader can earn profits. Through a laboratory experiment, our study verified that a linkage punishment leader actually appears under the leader support system, which makes it easier to achieve high cooperation in the PGG.

This idea of solving the second-order free rider problem under a linkage punishment leader originally was proposed in evolutionary simulation research conducted by Matsumoto and Jin (2010) [22]. They set up a game on a computer program comprising groups of one leader and the other followers. Each group

underwent a PGG stage, a support stage, and a punishment stage. The authors examined the strategy of the leader and follower that earned higher benefit and how this strategy evolved. Furthermore, the authors implemented a "dismissal mechanism" whereby the leader would be removed if s/he did not contribute to the increase of the group follower's profit. In such a situation, they showed that a linkage punishment leader evolved and that a high cooperation in the PGG and a high level of follower support was achieved. On the other hand, they found that without this dismissal mechanism, a self-focused punishment leader that prioritized the individual profit of the leader evolved, which resulted in low cooperation in the PGG.

Thus, our study incorporated Matsumoto and Jin's (2010) idea in a laboratory experiment, but a leader dismissal system was not set in the experiment. Simple logic predicts that a self-focused punishment leader may spread, but the leader in the experiment might potentially execute linkage punishment because of his/her prediction that long-term support is gained not only through self-focused punishment. Furthermore, we considered that the actual support from followers for a self-focused punishment leader is lower than that for a linkage punishment leader because followers under a self-focused punishment leader try to resist such a leader. As such, with the "dismissal" system as a psychological predication of the leader and the follower's actual

actions, a linkage punishment leader is predicted to arise even without a dismissal system.

1.4. Punishment by a leader

There have been a few laboratory experiments that investigated how setting one leader as a punisher affects the solution of the public goods problem [23, 24]. While these are pioneering studies on leader punishment, they are critically different from the leader support system. In these previous studies on leader punishment, the capital for punishment is provided externally and not from the support by followers. As discussed, an incentive for leaders to punish the first-and second-order free riders is provided because they receive support from followers. If they do not have any chance to obtain support from followers, leaders cannot obtain any material benefit by punishing followers, which significantly decreases the incentive for punishment. By comparing conditions with and without support, our study attempted to show that punishment by leaders is heavier if there is a system in which the leader receives support from followers, which makes it easier to achieve high cooperation in the PGG.

1.5. Experiment hypotheses

In our experiment, six-person groups were assembled, comprising five followers who

engaged in a PGG and one leader executing punishment. Thereafter, settings included whether capital for the leader's punishment came from followers' support (support-present condition) or from the outside (no-support condition). Four hypotheses were set based on these arguments.

Hypothesis 1: Punishment toward non-cooperators in the PGG is more likely to occur in a support-present condition than a no-support condition (H1a), making it likely that high cooperation is achieved (H1b).

Hypothesis 2: Within a support-present condition, the linkage punishment leader executes stronger punishment to PGG non-cooperators than the non-linkage punishment leaders do (H2a), making it easier to achieve high PGG cooperation (H2b).

Hypothesis 3: Within a support-present condition, the linkage punishment leader executes stronger punishment for non-supportive followers than the non-linkage punishment leaders do (H3a), making it easier to achieve a high level of support (H3b).

Hypothesis 4: As a result of Hypotheses 2 and 3, the profits of the linkage punishment leaders are higher than those of the non-linkage punishment leaders (H4a) and the profits of the followers are higher under the linkage punishment leader than those under the non-linkage punishment leader (H4b).

2. Methods

2.1. Participants

In total, 270 university students participated in this experiment, of which 162 students (27 groups) participated in the support-present condition and 108 students (18 groups) participated in the no-support condition. Participants were recruited via a university portal website, and monetary reward was emphasized during recruitment. For 15 other participants in the support-present condition, a six-person group could not be assembled because there were not enough members. These groups were handled by adding as participants staff members who did not know the experiment details, and so, these groups were excluded from the analysis.

2.2. Procedure

Eighteen participants participated in each session of the experiment. After reading explanations of PowerPoint slides, the participants answered confirmation tests that questioned their understanding of the experiment details. Neutral words were selected for explanation. After confirming that all participants understood the experiment details, they were allocated randomly to one of three six-person groups. After running a trial period once, the participants started the real session.

The details of the experimental transactions are as follows. First, at the beginning of the session, the roles of one leader who executes punishment and five followers who engage in the PGG were selected randomly. The participants were told that these roles and the composition of the group members would remain unchanged throughout the experiment. The transactions comprised three stages: a PGG stage, a support stage, and a punishment stage. The participants were told before the beginning of the experiment that these periods would be repeated 15 times, and that the tokens they earned during transactions would be redeemed as monetary remuneration.

PGG stage. Each of the six members, including the leader, was given 100 tokens at the beginning of the stage. The five followers decided whether to contribute all 100 tokens to the group pool or not at all. The tokens each follower contributed were doubled and distributed equally to five followers except for the leader. This meant that each time one follower made a contribution, all five followers received 40 tokens each. The leader was completely independent from the other followers. Although the leader was given 100 tokens, like the other followers, s/he did not make decisions during this stage and simply earned 100 tokens.

Support stage. In the support-present condition, an additional 20 tokens were provided to each of the six members, including the leader. The five followers other than the leader decided whether to provide (support) the 20 tokens to the leader or not. If a follower decided to support the leader, the follower lost the 20 tokens and a leader obtained the 20 tokens. There was nothing for the leader to decide.

In the no-support condition, the leader was given 120 tokens while the five followers were given 20 tokens each. There was nothing for any group follower or the leader to decide in this condition. The reason why only the leader was given 120 tokens is that is the maximum-attainable amount in the support-present condition if all five followers were to provide their 20 tokens to a leader. By setting the punishable amount to at least the same as the leader in the no-support condition as in the support-present condition in the subsequent punishment stage, it was ensured that there was no disadvantage for the no-support condition.

Punishment stage. The leader used the amount earned in the support stage as capital, that is, a fixed 120 tokens in the no-support condition; 20 + (the number of support followers) X 20 in the support-present condition (minimum 20, maximum 120). Then, the leader determined, in increments of 20 tokens, how much to reduce each follower's

tokens. The punishment rate was double, meaning that if a leader used 20 tokens to punish a certain follower, the follower would lose 40 tokens. As long as there was sufficient capital, the leader could reduce anyone's amount of tokens. The amount the leader did not use for punishment was added to the leader's own profit.

The PGG results, that is, who contributed or did not contribute to the group, were provided to all six members after the support stage in accordance with previous pool punishment system studies [18–20], in which the PGG result was unknown at the stage at which participants decided whether to bear the punishment cost to the pool. In addition, all six members were informed about followers who supported the leader after the support stage. Thus, during the punishment stage, the leader was able to decide whom to punish after ascertaining who contributed in the PGG and who supported him/her. Furthermore, all members were informed who had been punished and by how much immediately after the leader's decision. Therefore, followers could know the punishment type of their leader.

These three stages were repeated 15 times. Experiment control was conducted using Ztree [25]. The total attained score was converted to money using the rate 1 token ＝0.7 yen, and the converted amount was provided plus 500 yen (the show-up fee) given

to participants who concluded the experiment. The average remuneration amount was 2,117 yen.

3. Results

3.1 Support-present condition versus no-support condition.

First, the support-present condition and no-support condition were compared. The average total contribution of the five followers in the PGG and the leaders' average punitive amount for each non-contributor in the PGG through all 15 periods were calculated for each group. Figure 1 indicates this correlation. The punishment index did not use a simple average punitive amount because the number of non-contributors, who were the target of punishment, varied greatly depending on the group and period. Therefore, the leaders' average punitive amount for each non-contributor was calculated for each group by the number of non-contributors and set as the denominator, and the number of total punitive amounts the leader imposed upon the non-contributor was set as a numerator. The calculated amounts are used as the punishment index.

The leaders in the support-present condition punished more non-contributors than the leaders in the no-support condition did (Mann–Whitney U-test: $p = .001$). As for the

average group contribution in the PGG, there was no significance between conditions (Mann–Whitney U-test: p = .144). However, using the definition of "high cooperation level" as an average contribution of more than 80%, eight out of 27 groups (29.6%) in the support-present condition and none of 18 groups (0%) in the no-support condition achieved a high cooperation level, which was statistically significant (Fisher's exact test: p = .014). By setting the definition of "high cooperation level" to more than 70% and 60%, similar statistical significance was observed (p= .007 and p= .003, respectively). Thus, high PGG cooperation was achieved more easily for the support-present condition. It was not statistically significant with the U-test because polarization of high and low cooperation occurred in the support-present condition; there were several groups that resulted in low cooperation, even in the support-present condition, which weakened statistical power. Furthermore, there was a strong positive correlation between average punitive amount per non-contributor and the average total PGG contribution (r = .83, p < .001). These results supported Hypotheses 1a and 1b, which stated that punishment of the non-cooperators more likely occurred in the support-present condition than in the no-support condition, and thus, achieving high cooperation was also more likely to be achieved.

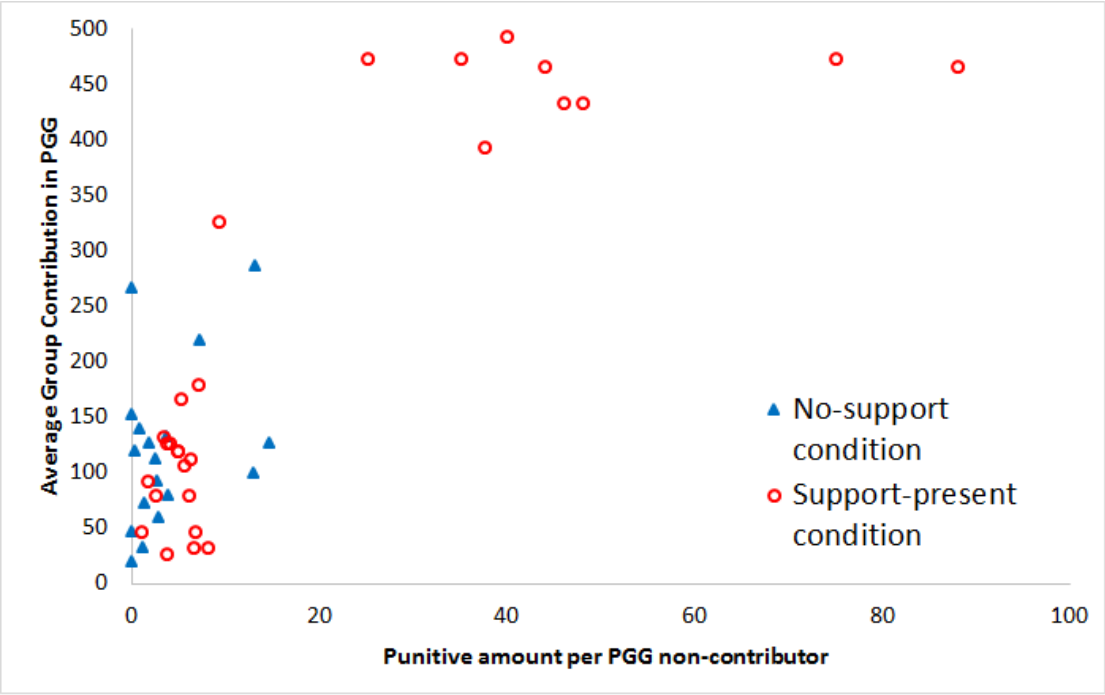

Fig. 1. Correlation between average punitive amount per PGG non-contributor and average group contribution in the PGG for 15 periods.

3.2. Comparing punishment types by leaders in the support-present condition.

Within the support-present condition, we analyzed what type of punishment by leaders achieved high cooperation. First, we calculated the amount of punishment that the leaders imposed on each of three follower's behavioral types—"non-contributing in PGG but supporting the leader," "contributing in PGG but non-supporting the leader," and "non-contributing in PGG and non-supporting the leader." Thereafter, leaders who had punished all three follower's behavioral types at least once were referred to as a linkage punishment (L) type, with other leaders set as a non-linkage (NL) type (leaders that

punished not all of three follower's behavioral types). There were 11 L-type and 16 NL-type leaders. One leader never encountered "non-contributing in PGG and non-supporting the leader" follower type throughout the 15 periods, but the leader punished the other two follower types. Thus, we regarded the leader as L-type. The average total PGG contribution, average total support amount, average punitive amount per non-contributor, average punitive amount per non-supporter, leader's total profit, and followers' total average profit per group were calculated for 14 periods but not the last 15th period. This was because all participants in our experiment knew the 15th period would be the last, and so, all the indexes tended to decline in the last period (see Figure 2). Thereafter, a Mann–Whitney U-test was conducted to determine whether a difference existed between L-type and NL-type leaders. The results showed that there was a significant difference in all categories, implying that the following hypotheses are supported: Hypothesis 2a and 2b (i.e., under an L-type leader, punishment toward a non-cooperator is more strongly imposed ($p < .001$), and thus, PGG cooperativeness is higher ($p = .001$)); Hypothesis 3a and 3b (under L-type leaders, punishment toward non-supporters is more strongly imposed ($p= .017$), and therefore, support toward the leader is higher ($p< .001$)); and Hypothesis 4a and 4b (as a result of these developments, profit of the L-type leader is higher ($p= .007$) and followers' profits under the L-type

leader are higher (p= .002)).

Next, punishment types were categorized in more detail. With leaders that punished all three followers' behavioral types at least once referred to as linkage punishment (L) types, leaders that punished followers that were *non-contributing but supporting* and *non-contributing and non-supporting* while never punishing *contributing and non-supporting* followers were referred to as group-focused punishment (G) types; leaders that punished only followers that were *contributing and non-supporting* and *non-contributing and non-supporting* while never punishing *non-contributing and supporting* followers were referred to as self-focused punishment (S) types. There were 11 L-type, 7 G-type, and 5 S-type leaders. Although three leaders did not fit any of the three categories, they were few in number and could not be interpreted easily. Thus, they were excluded from the analysis.

The total PGG contribution, total support amount to the leader, punitive amount per non-contributor, punitive amount per non-supporter and average profit of leader and followers were calculated for each period (see Figure 2). The punitive amount per non-contributor and punitive amount per non-supporter were impossible to calculate in some periods because there were no non-contributors or non-supporters in those periods. We excluded these data when calculating the averages.

The Mann–Whitney U-test was conducted to determine whether a difference existed between the leader types for the first 14 periods. Bonfirroni's correction was used to determine the significance of comparisons of the three leader types L, S, and G from this point onward. As for the average group contribution, the followers under L-type leaders contributed more to the PGG pool than under G-type leaders (p= .039) and S-type leaders (p= .048). The support for L-type leaders was higher than that for G-type leaders (p= .003) and marginally higher than that for S-type leaders (p= .081). Although the S-type leaders punished non-supporters similarly to L-type leaders, as we describe later in this subsection, the support amount of S-type leaders tended to be lower than that of L-type leaders.

The average punitive amount per non-contributor throughout the 14 periods was larger for L-type leaders than for G-type (p= .009) and S-type leaders (p= .003). Concerning the average punitive amount per non-supporter for the14 periods, L-type leaders punished more than G-type leaders did (p= .036), while there was no difference between L- and S-type leaders (p= .801).

Finally, the total profit of leaders and followers except for the last 15th period was analyzed. The Mann–Whitney U-test revealed that the leader's profit was higher for L-type leaders than for G-type leaders (p= .027) because G-type leaders could not attain

support and their profit remained low. There was no statistically significant difference between L- and S-type leaders (p= .540). This showed that S-type leaders reached a certain profit amount by attaining support. As for the followers, the Mann–Whitney U-test revealed that followers under L-type leaders had higher profits than those under S-type leaders (p= .039) and marginally higher profits than those under G-type leaders (p= .078), which resulted from achieving higher PGG cooperation under an L-type leader than under an S-type or a G –type leader. In addition, to examine the superiority of the leader within a group, the difference in profit between leaders and followers was calculated (L types 105; S types 614; G types 217) and analyzed using a U-test. The results showed that S-type leaders had higher profit differences than L-type (p = .003) and G-type leaders (p = .018) did, which means that the S-type leaders were likely to hold relatively dominant positions within the group.

Fig. 2. Average group contribution to the public good (A), average support amount for the leader (B), average punishment for one non-contributor in the PGG (C), average punishment for one non-supporter (D), average profit of leaders (E), and average profit of followers (F) over 15 periods of play under the linkage leader , group-focused leader , and self-focused leader.

4. Discussion

4.1. Differences between the leader support and non-support systems

Comparing the support-present condition and the no-support condition, high cooperation in the PGG was unlikely to be attained in the no-support condition because of weak punishment. These results are sufficiently understandable because the leader has no material incentive to punish in the no-support condition. Baldassarri et al. (2010) [23] showed that cooperation was achieved under a leader in spite of no support by followers in their experiment. At first glance, this may seem contradictory to our results. However, the participants in their study included acquaintances, and so, punishment may have been executed in consideration of future reputation [23, pp. 11026]. In a completely anonymous situation like our experiment, it was difficult to solve the public goods problem by a leader that was not supported by his or her followers.

4.2. How punishment type affects behavior and profits of group and leader.

We analyzed how the behavior and profits of group and leader differed according to the leader's punishment type within a support-present condition. The results showed that an L-type leader strongly punished PGG non-cooperators—the first-order free riders—making it easier to achieve high cooperation. Furthermore, L-type leaders strongly punished non-supporters—the second-order free riders—making a high

support level easy for the leader to maintain punishment. In addition, compared to other punishment types, both L-type leaders themselves and their followers earned higher profits. In previous studies, the pool punishment system with the second-order punishment has been shown to derive stable group cooperation [18, 19], but the reasons for such a punishment rule being generated have not been discussed. In our study, by perceiving the punishment executer to be human, we showed that such a rule could lead to an increase in the profit of the punishment executers themselves. This finding makes a significant contribution to understanding the emergence of the pool punishment system with second-order punishment.

Support for a G-type leader that only punished non-PGG cooperators—only the first-order free riders—was low. Because the leader did not have much capital, s/he could not strongly punish the first-order free riders. Thus, it was hard to accomplish cooperation in the PGG. The group-focused punishment is an action that focuses more on the profit of the entire group rather than the profit of the leader himself/herself. Simple intuition might dictate that this is a desirable action for a group. However, it is ironic that this induces a lack of support from followers, resulting in a state in which both the leader and followers cannot attain much profit. It is interesting that an L-type leader who sets a high value on not only profit of the group but also his/her own profit

increases the profit of the entire group as a result.

High PGG cooperation is not achieved under an S-type leader who punishes only non-supporters—only the second-order free riders. This is because an S-type leader is not beneficial for the group at all. Therefore, the fear of punishment was the only incentive for followers to support the leader. The fact that the support of the S-type leaders were lower than that of the L-type leaders reflects the antagonism of the followers, which implies that the L-type punishment is a superior method that can gain and maintain support from the group. However, there was no significant difference between the S-type and L-type leaders in terms of the total profit, as the former was able to attain substantial profit. Moreover, under the S-type leader, the profit difference between the followers and leader was the largest, showing that the S-type leader could settle in a comparatively superior position within the group. These results showed that it was reasonable for the leader in our experiment to select self-focused punishment, that is, the leader support system risks causing tyranny by invoking the fear of the leader.

4.3. Strong and weak points of the leader support system

Our study showed that the second-order free rider problem would be solved if an L-type leader appeared under a leader support system. In addition, the leader support system

has the effect of suppressing retaliation or anti-social punishment observed in peer punishment [26-28]. Under the leader support system, the provision of resources to the leader occurs in a concentrated manner, causing skewed distribution in punishment capability. Therefore, even if followers are given a chance to retaliate against their leader, followers will hesitate to execute retaliation against the leader because they predict a strong retaliation in return from him/her. On the other hand, because the individual differences in the potential punishment capability are not so large in peer punishment, retaliatory action is more likely to occur. As a result, the leader support system can maintain unilateral punishment more easily.

The leader support system, however, has a serious problem. Unilateral punishment results in good consequences for the group only under L-type leaders. Under the other leaders, in particular, under S-type leaders, the state of the group may worsen. In our experiment, the followers under S-type leaders were non-cooperation with the PGG but continued supporting the leader, which generated no profit for the followers. As a result, a distorted state arose, in which the profit of the followers was low, but that of the leader was high. This risk of tyranny should be higher in actual societies than in a laboratory experiment as societies have larger group sizes than the experiment did. As group size increases, the inequality in punishment capability between the leader and

followers increases, potentially creating a state in which followers cannot resist the leader's tyranny even if they want to. Furthermore, relationships last longer in an actual society compared to a laboratory. With the metamorphic effects of power [29], people become selfish by continuously exercising power. If this experiment lasted over the longer term, a threat develops in which even if a leader was an L-type leader in the beginning, s/he may gradually become an S-type leader in the long term. In this manner, such factors as an increase in group size and a long-lasting administration act more strongly in an actual society than in a laboratory, heightening the risk of tyranny. Acemoglu and Robinson (2012) [30] pointed out the importance of governance by law and a democratic system in order to prevent leader tyranny. They argued this with many examples that highlight how democratic insufficiency resulted in exploitation by a leader. Our experiment provides a methodology to approach this issue empirically.

4.4. Future research

In the future, there is a need to explore how leader support systems emerge. In our research, a system was one in which followers supported a leader and only this leader could punish followers. However, to examine how such a system emerges, it is necessary to set a state in which the leader is not present in the beginning of the experiment. To be specific, three stages should be set, that is, a PGG stage in which all members

participate, a support stage in which anyone can support anyone, and a punishment stage in which anyone can punish anyone. In the experiment, it would be possible to examine how a high level of group cooperation is achieved as a result of concentrated support for a specific individual who spontaneously executes a linkage punishment. If everyone is completely equal and does not receive any information, concentrated support for a specific individual is less likely to occur. In an actual society, concentrated support for a specific individual who should become a leader starts after the reputation of the leader spreads concerning his/her qualifications and charisma. The authority of this leader solidifies by him/her taking actions to meet followers' expectations. Even in the laboratory, governance by a specific individual might naturally arise by sharing the qualities of each individual. Empirically addressing the natural rise of governance can be seen as equivalent to reproducing the creation of "Leviathan" [31] within human society in a laboratory. This would be an ambitious attempt, and such future research is anticipated eagerly.

Ethics. The Waseda University Ethical Review Board specifically approved this study. Written informed consent was obtained from all participants prior to beginning the experiment.

Data accessibility. The datasets supporting this article have been uploaded as part of the supplementary material.

Authors' contributions. H.O. & N.J. designed research and H.O. & K. S. performed the experiment; H. O & M.W. analyzed the data and wrote the paper; all authors interpreted the results and reviewed and approved the manuscript.

Competing interests. The authors declare no competing interests.

Funding. This study was funded by JSPS KAKENHI Grant Number 25885057.

References

1. Hardin, G. The Tragedy of the Commons. *Science* **162**, 1243 (1968). (doi: 10.1126/science.162.3859.1243)
2. Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press, Cambridge, 1965)

3. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev* **90**, 980–994 (2000). (doi:10.1257/aer.90.4.980)

4. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002). (doi:10.1038/415137a)

5. Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: Self governance is possible. *Am. Pol. Sci. Rev* **86**, 404-417 (1992). (doi:10.2307/1964229)

6. Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M.A. Positive interactions promote public cooperation. *Science* **325**, 1272–1275 (2009). (doi:10.1126/ science.1177418)

7. Boyd, R. & Richerson, P. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobio* **13**, 171-195 (1992). (doi: 10.1016/0162-3095(92)90032-Y)

8. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499-502 (2004). (doi:10.1038/nature02978)

9. Cinyabuguma, M., Page, T. & Putterman, L. Can second-order punishment deter perverse punishment? *Experimental Economics*, **9**, 265-279 (2006). (doi:10.1007/s10683-006-9127-z)

10. Boyd. R., Gintis. H., Bowles, S. & Richerson, P. J. The evolution of altruistic

punishment. *Proc Natl Acad Sci USA* **100**, 3531–3535 (2003). (doi:10.1073/pnas.0630443100)

11. dos Santos, M., Rankin, D. J. & Wedekind, C. The evolution of punishment through reputation. *Proc. R. Soc. Lond. B* **278**, 371–77 (2011). (doi:10.1098/rspb.2010.1275)
12. Barclay, P. Reputational benefits for altruistic punishment. *Evo. Hum. Beh.* **27**, 344 (2006). (doi:10.1016/j.evolhumbehav.2006.01.003)
13. Pinker, S. *The false allure of group selection*. Retrieved July 20, 2012, from http,//edge.org/conversation/the-false-allure-of-group-selection.
14. Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends. Ecol. Evol.* **30**, 98–103 (2015). (doi:10.1016/j.tree.2014.12.003)
15. Guala, F. Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35**, 1-59 (2012). (doi:10.1017/S0140525X11000069)
16. Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol* **51**, 110–116 (1986). (doi:10.1037/0022-3514.51.1.110)
17. Yamagishi, T. The provision of a sanctioning system in the United States and Japan. *Social.Psycho. Quarterly* **51**, 265–271 (1988). (doi:10.2307/278692)
18. Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. Social learning promotes

institutions for governing the commons. *Nature* **466**, 861–863 (2010). (doi:10.1038/nature09203)

19. Perc, M. Sustainable institutionalized punishment requires elimination of second-order free-riders. *Scient. Rep.* **2**, 344 (2012). (doi:10.1038/srep00344)

20. Traulsen, A., R¨ohl, T. & Milinski, M. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B.* **279**, 3716–3721 (2012). (doi:10.1098/rspb.2012.0937)

21. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, Cambridge, 1990).

22. Matsumoto, Y. & Jin, N. Co-evolution of leader traits and member traits in social dilemmas (in Japanese). *Japanese J. Exp. Soc. Psycho.* **50**, 15-27 (2010). (doi: 10.2130/jjesp.50.15)

23. Baldassarri, D. & Grossman, G. Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11023–11027 (2011). (doi:10.1073/pnas.1105456108)

24. O'Gorman, R., Henrich, J. & Van Vugt, M. Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proc. Biol. Sci.* **276**,323–329 (2009). (doi:10.1098/rspb.2008.1082)

25. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* **10**, 17-178 (2007). (doi:10.1007/s10683-006-9159-4)

26. Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really govern ourselves? *J. Public Econ.* **92**, 91-112 (2008). (doi:10.1016/j.jpubeco.2007.04.008)

27. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**,1362–1367 (2008). (doi:10.1126/science.1153808)

28. Cinyabuguma, M., Page, T.& Putterman, L. Cooperation under the threat of expulsion in a public goods experiment. *J. Pub. Econ.* **89**, 1421–1435 (2006). (doi:10.1016/j.jpubeco.2004.05.011)

29. Kipnis, D. Does power corrupt? *J Pers Soc Psycho* **24**, 33-41 (1972). (doi:10.1037/h0033390)

30. Acemoglu, D. & Robinson, J. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty* (Crown Business, New York, 2012)

31. Hobbs, T. *Leviathan: Or the Matter, Forme, and Power of a Common-Wealth Ecclesiasticall and Civill*, ed. by Ian Shapiro (Yale University Press; 2010) (1651).

Supplementary Information

Instruction of the experiment

After a brief verbal introduction, participants read the following instructions on the computer monitor telling them that they would take part in an experiment on decision making.

General Guidance

This is an experiment about decision making. You will be paid for participating, and the amount of money you will earn depends on the decisions that you and the other participants make. At the end of today's session you will be paid in cash for your decisions privately.

You will never be asked to reveal your identity to anyone during the course of the experiment. Your name will never be associated with any of your decisions.

At this time, you will be given 500 yens (=5~6 dollars) for coming on time. All the money that you earn after this experiment will be yours to keep.

Earnings

In this experiment you are in a group of size 6 (you plus 5 others) and you will be asked to make a series of choices about how to allocate a set of tokens. You and the other subjects has been randomly assigned to the group, and you *will not* be able to know each other's identities. But the group members remained the same throughout the experiment.

The details of the experimental transactions are as follows. There are two different roles in the experiment. Five members named A, B, C, D and E will play the same role, but one member named Z will play a different role. Who will be assigned as Z will be selected randomly in the beginning of the experiment and these roles remained the same throughout the experiment. The experiment comprised three stages, 1st stage, 2nd stage and 3rd stage. These stages will be repeated 15 times, and the tokens you earn during transactions will be redeemed as monetary remuneration.

Now, let us explain the details of each stage.

1st stage:

Each of the six members, including Z, are given 100 tokens at the beginning of the stage. The

members except for Z are asked to decide whether to contribute all 100 tokens to the group pool or not at all. The tokens each member contributed are doubled and distributed equally to five members except for Z. This means that each time one member make a contribution, all five members except for Z received 40 tokens each. Z was completely independent from the other members. Although Z are given 100 tokens, like the other members, s/he does not make decisions during this stage and simply earns 100 tokens.

-Examples of choices you will make in this experiment and earnings

Example 1: Suppose that you are A, not Z. You and the other 4 members all contribute 100 tokens to a pool. You will earn:
100 (initial endowment) - 100 (the tokens you gave)
+ 0.4 * 500 (the sum of tokens 5 members gave)
=200
Example 2: Suppose that you are B, not Z. You and the other 4 members all contribute nothing. You will earn:
100 (initial endowment) - 0 (the tokens you gave)
+ 0.4 * 0 (the sum of tokens 5 members gave)
=100
Example 3: Suppose that you are B, not Z. You contribute nothing and all the other members contribute 100 tokens each. You will earn:
100 (initial endowment) - 0 (the tokens you gave)
+ 0.4 * 400(the sum of tokens 5 members gave)
=260.
Example 4: Suppose that you are Z. You do not make any decision. You will earn:
100 (initial endowment) .

2nd stage (support-present condition):

An additional 20 tokens are provided to each of the six members, including Z. The five members other than Z decide whether to provide the 20 tokens for Z or not. If a member decides to provide his/her tokens for Z, s/he loses the 20 tokens and Z obtains the 20 tokens. There are nothing for Z to decide.

-Examples of choices you will make in this experiment and earnings

Example 1: Suppose that you are A, not Z. You provide 20 tokens for Z. You will earn:
20(initial endowment) - 20 (the tokens you provide)

=0

Example 2: Suppose that you are B, not Z. You provide nothing for Z. You will earn:

20 (initial endowment) - 0 (the tokens you gave)

=20

Example 3: Suppose that you are Z. You do not make any decision. A, B, C,D and E provide 20, 0, 20, 20, 0 to you respectively. You will earn:

20(initial endowment) +60 (the tokens you are provided by the other members)

=80

2nd stage (no-support condition):

Z are given 120 tokens while the other five members are given 20 tokens each. There is nothing for any group members to decide in this stage

3rd stage:

Z can use the amount earned in the 2nd stage as capital, that is, the fixed 120 tokens (in the no-support condition) / 20 + (the number of members who provided their tokens) X 20(in the support-present condition). Then, Z determines, in increments of 20 tokens, how much to reduce each member's tokens. If Z uses 20 tokens to reduce the token of a certain member, the member will lose 40 tokens. As long as there is sufficient capital, Z can reduce anyone's amount of tokens. The amount Z does not use for reduction is added to Z's own profit.

-Examples of choices you will make in this experiment and earnings

Example 1: Suppose that you are A, not Z. you got 200 tokens in 1st stage and 20 tokens in 2nd stage. Z decides to reduce 40 tokens from you. You will earn:

200 (1st stage earning) + 20 (2nd stage earning) – 40 (the reduction by Z)

=180 (the total earning in the period)

Example 3: Suppose that you are Z. you got 100 tokens in 1st stage and 80 tokens in 2nd stage. You decides to use 60 tokens in total to reduce the other members' tokens. You will earn:

100 (1st stage earning) + 80 (2nd stage earning) – 60 (the use to reduce the other members' tokens)

=120 (the total earning in the period)

Feedbacks:

All six members are informed about the results of 1st stage, that is, who contributes or does not contribute to the group, after the 2nd stage,. In addition, all the members are informed about

members who provide their 20 tokens for Z after the 2nd stage as well. Thus, during the 3rd stage, Z is able to decide whose tokens to reduce after ascertaining who contributed in the 1st stage and who provided their tokens for Z. Furthermore, all members are informed whose tokens were reduced and by how much immediately after Z's decision.

These three stages will be repeated 15 times. The total attained score will be converted to money using the rate 1 token＝0.7 yen, and the converted amount will be provided plus 500 yen (the show-up fee) given to you in the end of this experiment.

After this general instruction above, all participants would start the experiment after filling out a confirmation test.

Confirmation Test

Before you start to make your decision, we should solve all questions on the paper. Read carefully through the provided information and write down the number of points on the paper. We will watch you solving the examples, check whether you get the right answers and help you in case that there is a problem or a question.

Before the decision-making

Good, now everybody has correctly solved the problems. If anybody has any more questions raise your hand now. Otherwise let's practice how to make your decision on your computer screen.

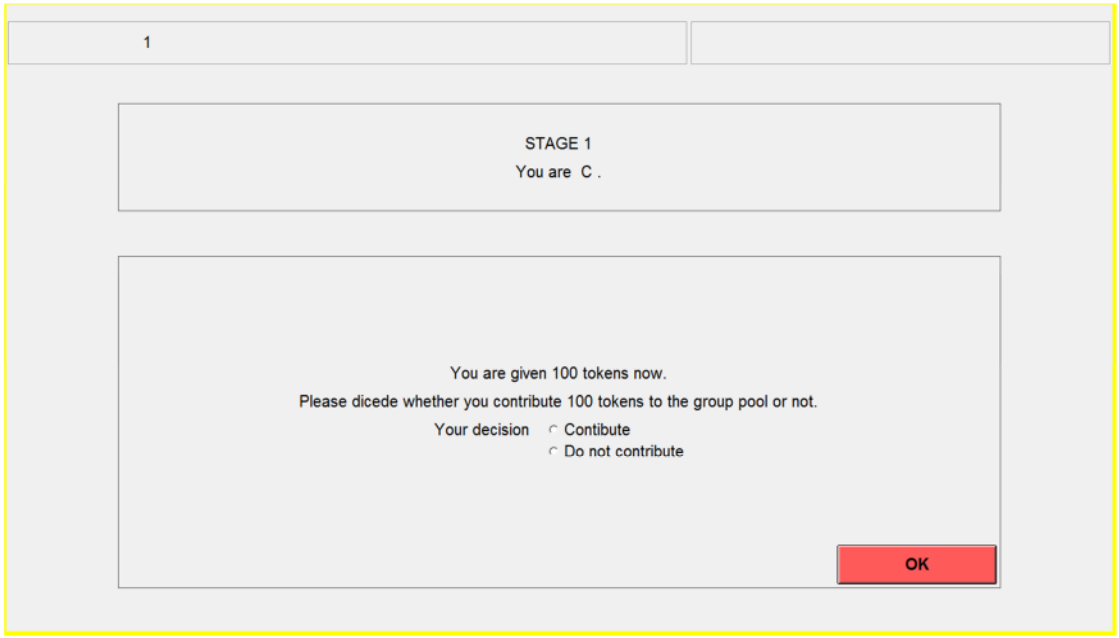

Figure 1. Screen shot of computer display when A, B, C,D,E make decisions in 1st stage.

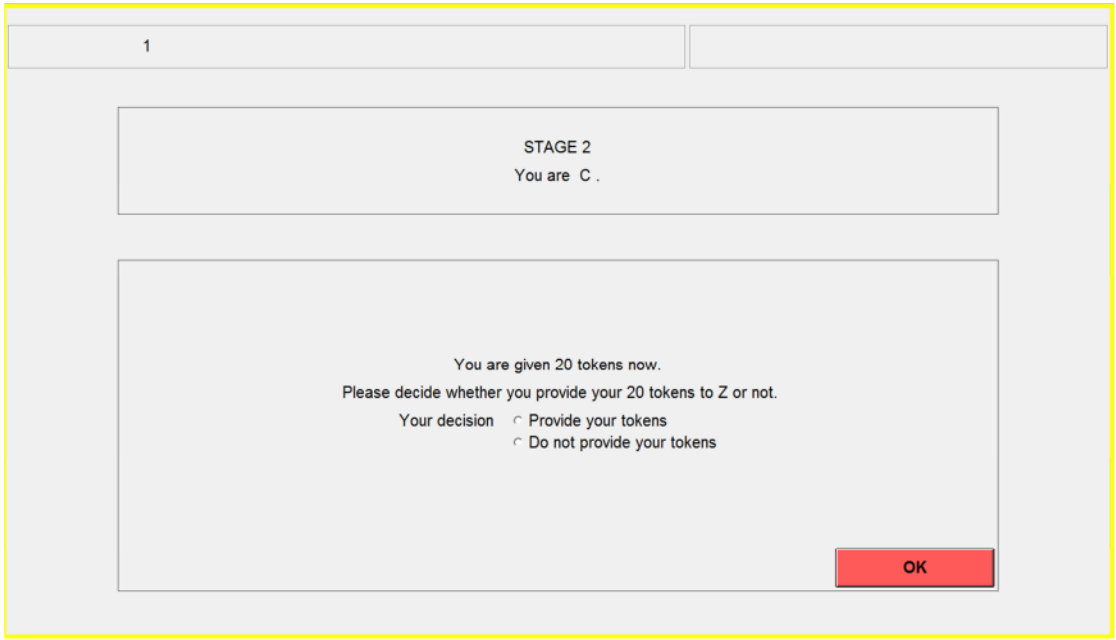

Figure 2. Screen shot of computer display when A, B, C,D,E make decisions in 2nd stage.

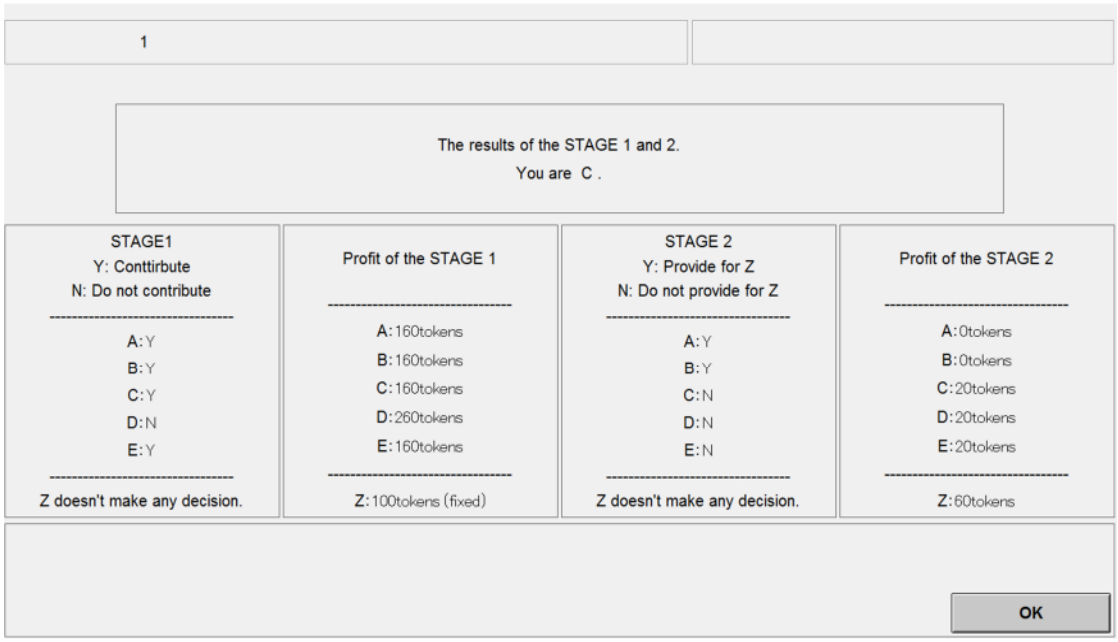

Figure 3. Screen shot of computer display when showing feedback after 2nd stage.

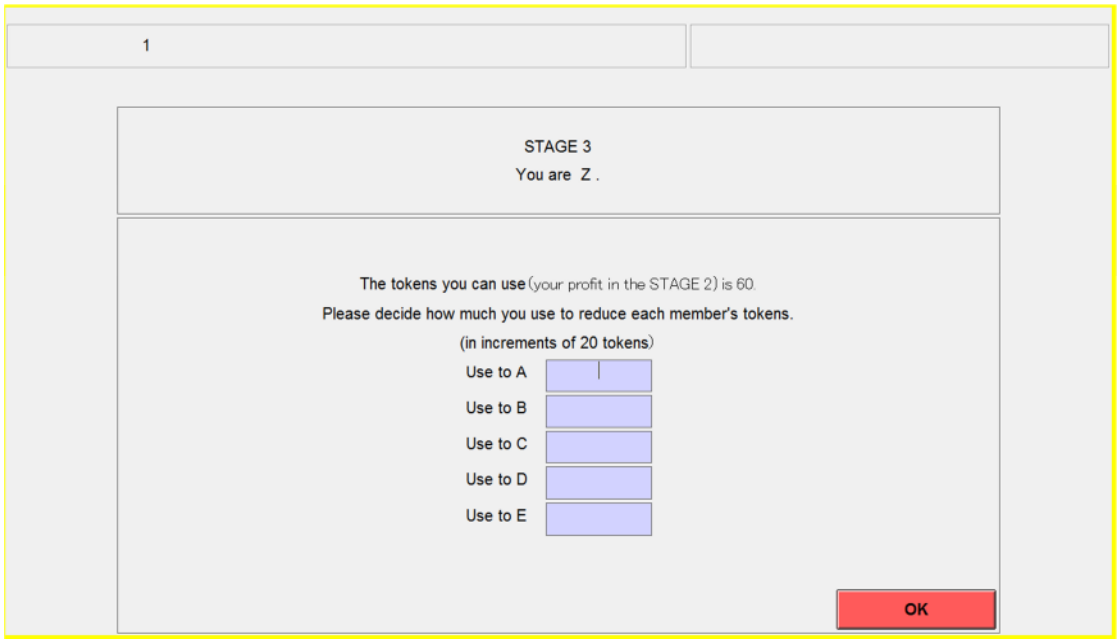

Figure 4. Screen shot of computer display when Z make decisions in 3rd stage.

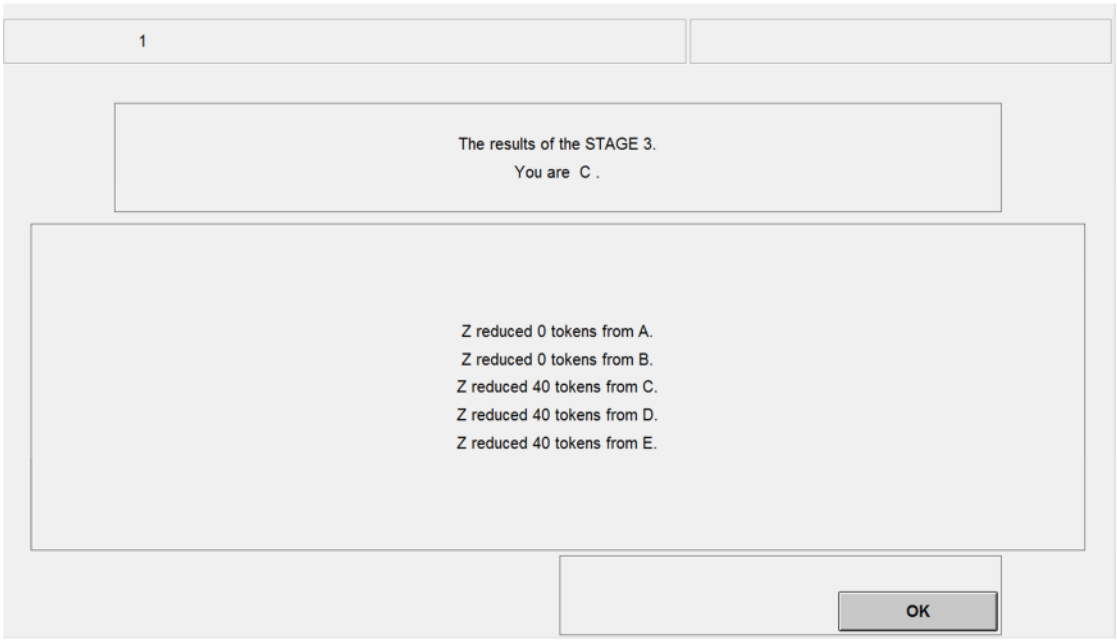

Figure 5. Screen shot of computer display when showing feedback after 3rd stage.