Accepted Manuscript

The Singapore Economic Review

Article Title:	The optimal incentive in promoting cooperation: punish the worst and do not only reward the best
Author(s):	Shuting Huang, Yali Dong, Boyu Zhang
DOI:	10.1142/S0217590820500071
Received:	18 April 2019
Accepted:	12 August 2019
To be cited as:	Shuting Huang, Yali Dong, Boyu Zhang, The optimal incentive in promoting cooperation: punish the worst and do not only reward the best, *The Singapore Economic Review*, doi: 10.1142/S0217590820500071
Link to final version:	https://doi.org/10.1142/S0217590820500071

World Scientific
www.worldscientific.com

The optimal incentive in promoting cooperation: punish the worst and do not only reward the best

Abstract

Incentive institutions that reward cooperators and punish free-riders are often used to promote cooperation in public goods games. We show that for incentives of intermediate size, a sanctioning institution that punishes the worst players can sustain full cooperation and that a rewarding institution can promote cooperation only if lower contributors also have the chance to win the reward. Furthermore, if the incentive institution can provide both reward and punishment, then it should use reward as much as possible. The group welfare is maximized when the punishment is just barely larger than the minimum required to obtain the full contribution.

Key words

Public goods game; cooperation; reward; punishment

JEL codes

C73, D02

1 Introduction

Sustaining cooperation in social dilemma games, such as public goods games (PGGs), is a fundamental challenge in economics and management science. This problem can be solved by establishing incentive institutions that reward cooperators and punish free-riders (Ostrom, 2005). Several types of institutions have been proposed to promote cooperation in PGGs. One type is characterized as an absolute incentive institution, where the institution punishes (or reward) all individuals whose contribution is less (or higher) than a predefined threshold (Sigmund et al., 2010; Sasaki et al., 2012; Traulsen et al., 2012; Zhang et al., 2014; Chen et al., 2015; Dong et al., 2019). A similar institution is that the reward (or punishment) amount increases (decreases) with the absolute contribution, see e.g., Galbiati, Vertova, (2008) and Putterman et al. (2011). Both theoretical and empirical studies indicated that absolute punishment can eliminate extremely selfish behaviors in a cooperative population (Sigmund et al., 2010; Traulsen et al., 2012; Zhang et al., 2014). In contrast, absolute reward is relatively ineffective in moving the equilibrium from the selfish one to the cooperative one (Sasaki et al., 2012; Chen et al., 2015). Another type is characterized as a relative incentive institution, where individuals who contribute an amount lower than the group average are more likely to be punished, and those who contribute higher than the group average are more likely to be rewarded (Yamagishi, 1986; Andreoni, Gee, 2012; Cressman et al., 2012; Cressman et al., 2013; Qin, Wang, 2013; Kamijo et al., 2014; Wu et al., 2014; Dong et al., 2016). A similar institution would be that the reward (or punishment) amount increases (decreases) with the relative contribution, see e.g., Falkinger et al. (2000). For relative punishment, a full contribution becomes a Nash equilibrium if the institution punishes the lowest contributor such that his or her payoff is slightly lower than that of the second lowest contributor (Andreoni, Gee, 2012). In contrast, there is much debate on the effectiveness of relative rewards, and the promotion of cooperation has been rarely observed in laboratory experiments (Cressman et al., 2012; Cressman et al., 2013; Wu et al., 2014; Dong et al., 2016).

The use of relative incentives is a common feature in many parts of human society such as in businesses, government institutions, schools, and competitive sports (Morgan, 2000; Andreoni, Gee, 2012). Currently, it is unclear what kind of relative rewards can promote cooperation. Furthermore, if the incentive institution can provide both reward and punishment, is it better to use more reward or punishment? In this paper, we consider three types of relative incentives, namely institutional reward (IR), institutional punishment (IP), and a mixture of reward and punishment (IRP). In line with previous studies, the institution rewards or punishes one player according to their relative contributions (Yamagishi, 1986; Andreoni, Gee, 2012; Cressman et al., 2012; Cressman et al., 2013; Qin, Wang, 2013; Kamijo et al., 2014; Wu et al., 2014; Dong et al., 2016). Specifically, we assume that the probability that a player is rewarded or punished has the form of the Tullock contest function. The Tullock contest function has been commonly used in the rent-seeking game. In this game, n players compete for a prize, and player i wins the prize with probability $P(\mathbf{x}) = x_i^s / \sum_{j=1}^n x_j^s$, where x_i is the effort of player i (Tullock, 1980; Hehenkamp et al., 2004; Chowdhury, Sheremeta, 2011; Ewerhart, 2015). In PGGs with institutional incentives, the prize can be seen as the reward, and the effort can be seen as the contribution level. The Tullock contest function has a free parameter s, which measures the probability that a player is rewarded or punished. At $s = 0$, all players except free-riders (or except full contributors) are equally likely to be rewarded (or punished). As s approaches infinity, the institution only rewards the highest contributor and punishes the lowest contributor. Most previous studies have considered relative incentives with specific s, e.g., Cressman et al. (2012), Cressman et al. (2013), Wu et al. (2014), and Dong et al. (2016) considered IR, IP, and IRP with $s = 1$, and Yamagishi (1986), Andreoni, Gee (2012), and Kamijo et al. (2014) considered IP with $s = \infty$. So far, the relation between s and the effectiveness of the incentives is still unknown.

The main purpose of this paper is to find the optimal relative incentive that is both effective in promoting cooperation and preserving group welfare. We do this in two steps. The first step is to calculate the s that optimizes the group contribution at the evolutionarily stable Nash equilibrium (NE) for fixed amounts of reward and punishment. The second step is to determine the amounts of reward and punishment that maximize the group welfare. The rest of this paper is organized as follows. Section 2 introduces the three types of incentives IR, IP, and IRP. Section 3 analyzes NEs and their evolutionary stabilities for PGGs with IR, IP, and IRP (related proofs are shown in the Appendices 1-3). Theoretical analysis shows that s should not be too large in IR so that lower contributors should also have the chance to be rewarded. In contrast, s should be as large as possible in IP, i.e., the institution should only punish the lowest contributor. Furthermore, when the incentive institution can provide both reward and punishment, the group welfare reaches maximum when the punishment is just barely larger than the minimum required to obtain full contribution. Section 4 discusses the main results.

2 Public goods game with institutional incentives

Consider a PGG with n players, where each has an initial endowment E. Each player decides how much of his endowment to contribute to a common pool. The total contributions to the common pool are multiplied by a factor r and split evenly among all n players. Suppose that player i contributes x_i to the common pool ($0 \leq x_i \leq E$). Then his payoff can be written as

$$f(x_i, \mathbf{x_{-i}}) = E - x_i + \frac{r}{n} \sum_{j=1}^{n} x_j, \tag{1}$$

where vector $\mathbf{x_{-i}}$ represents the contribution of the other $n-1$ players. For $1 < r < n$, $f(x_i; \mathbf{x_{-i}})$ is a decreasing function of x_i. In this case, the less the player contributes, the more the player receives. Thus, free-riding is the unique NE, although full contribution is better for the group.

Let us now introduce the (relative) incentive institution. In line with previous studies, the institution chooses to reward and/or punish one of the n players (Yamagishi, 1986; Andreoni, Gee, 2012; Cressman et al., 2012; Qin, Wang, 2013; Kamijo et al., 2014; Wu et al., 2014; Dong et al., 2016). The probability that a particular subject is rewarded or punished has the form of the Tullock contest function, i.e., player i wins the reward with probability

$$P_{IR}(x_i, \mathbf{x_{-i}}) = \frac{x_i^{s_R}}{\sum_{j=1}^n x_j^{s_R}}, \tag{2}$$

and is punished with probability

$$P_{IP}(x_i, \mathbf{x_{-i}}) = \frac{(E-x_i)^{s_P}}{\sum_{j=1}^n (E-x_j)^{s_P}}. \tag{3}$$

Thus, P_{IR} is increasing with the contribution and P_{IP} is decreasing with the contribution. In addition, no player deserves to be rewarded if they all contribute 0, and no one should be punished if they all contribute E.

We consider the reward amount to be R and the punishment amount to be P. If $0 < \sum_{j=1}^n x_j < nE$ (i.e., not all players contribute 0 or E), then the expected payoff for player i is

$$f(x_i, \mathbf{x_{-i}}) = E - x_i + \frac{r}{n}\sum_{j=1}^n x_j + \frac{R x_i^{s_R}}{\sum_{j=1}^n x_j^{s_R}} - \frac{P(E-x_i)^{s_P}}{\sum_{j=1}^n (E-x_j)^{s_P}}. \tag{4}$$

If $\sum_{j=1}^n x_j = 0$, then no one is rewarded, and the expected payoff for each player is $E - P/n$. Finally, if $\sum_{j=1}^n x_j = nE$, then no one is punished, and the expected payoff for each player is $Er + R/n$.

3 Results

We calculate symmetric NEs for PGGs with IR, IP, and IRP, in which all players contribute the same. Specifically, an interior state $0 < x^* < E$ is a NE if and only if $f(x, x^*) \leq f(x^*, x^*)$ for all $x \in [0, E]$, where

$$\begin{aligned} f(x, x^*) \quad = \quad & E - x + \frac{r}{n}(x + (n-1)x^*) \\ & + \frac{R x^{s_R}}{x^{s_R} + (n-1)x^{*s_R}} - \frac{P(E-x)^{s_P}}{(E-x)^{s_P} + (n-1)(E-x^*)^{s_P}}. \end{aligned} \tag{5}$$

is the payoff for a player who deviates from the NE strategy x^*.

In addition, we analyze the evolutionary stabilities of NEs by adaptive dynamics (Dieckmann, Law, 1996; Hofbauer, Sigmund, 1998; Dong et al., 2015). In the framework of adaptive dynamics, populations are assumed to be homogeneous, and the average contribution moves towards the direction in which mutants have the higher invasion payoff. Thus, a stable state of the adaptive dynamics can prevent the invasion of local mutations. The relationship between NE and fixed points of adaptive dynamics is well known: an interior NE must be a fixed point, but a fixed point need not be a NE (Hofbauer, Sigmund, 1998).

Consider a homogeneous population with contribution x. The adaptive dynamics for Eq.(5) is written as

$$\frac{dx}{dt} = \frac{\partial f(y,x)}{\partial y}|_{y=x} = -1 + \frac{r}{n} + \frac{Rs_R(n-1)}{n^2 x} + \frac{Ps_P(n-1)}{n^2(E-x)}. \tag{6}$$

where $f(y,x)$ is the payoff for the mutant. Since $x \in [0,E]$, we further add two boundary conditions. The free-riding state $x = 0$ is a stable fixed point if $dx/dt < 0$ as $x \to 0$, and the cooperative state $x = E$ is a stable fixed point if $dx/dt > 0$ as $x \to E$ [5].

Proposition 1: the reward case $R > 0$ and $P = 0$
If $R \geq (n-r)E$, then the cooperative state $x^* = E$ is the unique NE for $s_R \geq n/(n-1)$. If $R < (n-r)E$, then $x^* = Rs_R(n-1)/n(n-r)$ is the unique NE only if $s_R \leq n/(n-1)$. Otherwise, the game has no symmetric NE. Furthermore, a NE in IR must be globally stable under adaptive dynamics.

Proportion 1 indicates that in IR, the optimal s_R for promoting cooperation depends on the incentive size R. For $R \geq (n-r)E$, a larger s_R can help to sustain cooperation. In this case, the cooperative state is a NE and is globally stable under adaptive dynamics (see Figure 1b). However, for $R < (n-r)E$, s_R larger than $n/(n-1)$ is detrimental to stable contribution. Notice that the contribution at the NE $x^* = Rs_R(n-1)/n(n-r)$ is increasing in s_R, the optimal s_R should be $n/(n-1)$, and the corresponding x^* is $R/(n-r)$ (see Figure 1a).

Proposition 2: the punishment case $R = 0$ and $P > 0$
If $P \geq (n-r)E/n$, then the cooperative state $x^* = E$ is a NE, and it is the unique NE for $s_P \geq n(n-r)E/P(n-1)$. If $P \leq (n-r)E$, then both $x^* = 0$ and $x^* = E - Ps_P(n-1)/n(n-r)$ can be a NE. In particular, $x^* = 0$ is a NE only if $s_P \leq n(n-r)E/P(n-1)$, and $x^* = E - Ps_P(n-1)/n(n-r)$ is a NE only if $n/(n-2) \leq s_P < n(n-r)/EP(n-1)$. Otherwise, the game has no symmetric NE. Furthermore, both the cooperative and the free-riding NEs are stable (if exist), and the interior NE is unstable.

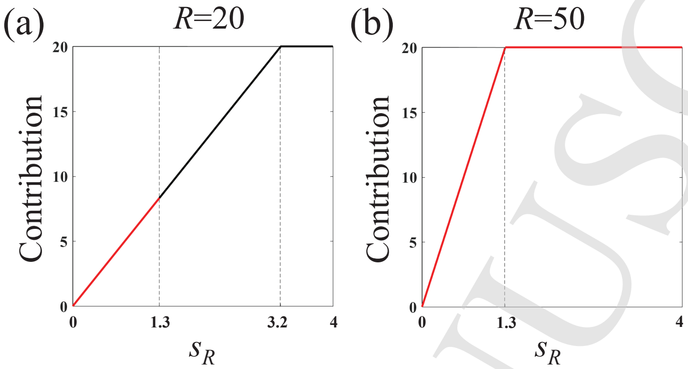

Figure 1: Institutional reward. Parameters are taken as $E = 20$, $n = 4$, $r = 1.6$. NE are denoted by red lines and stable equilibria of Eq.(6) are denoted by solid lines. **(a)** $R = 20$. The optimal $s_R = 1.3$. **(b)** $R = 50$. The cooperative state is the only stable NE for $s_R > 1.3$.

Proposition 2 points out a larger s_P is always beneficial to cooperation (see Figure 2). When $s_P > n(n-r)E/P(n-1)$, the free-riding state is no longer a NE, and the cooperative state is the only stable NE for $P \geq (n-r)E/n$.

Proposition 3: a mixture of reward and punishment $R > 0$ **and** $P > 0$
IRP has at most three NEs, a cooperative NE and two interior NEs. The NE condition for $x^* = E$ *is independent of* s_P*. If* $P \geq (n-r)E/n$*, then the cooperative state* $x^* = E$ *is a NE for all* s_R*. If* $P + R/n \geq (n-r)E/n$*, then the cooperative state* $x^* = E$ *is a NE for* $s_R \to \infty$*. If* $P+R/n < (n-r)E/n$*, then the cooperative state* $x^* = E$ *cannot be a NE. In particular, the existence of an interior NE is impossible when* $s_R \geq n(n-r)E/R(n-1)$ *or* $s_P \geq n(n-r)E/P(n-1)$*. Furthermore, the cooperative NE must be stable, and at most one interior NE is stable.*

Proposition 3 indicates that if $P + R/n \geq (n-r)E/n$, then it is better to reward the highest and punish the lowest. In this case, the cooperative state is the only NE (see Figure 3b). However, if $P + R/n < (n-r)E/n$, the cooperative state cannot be a NE. In this case, IRP may have one stable interior NE for small or intermediate s_R and s_P, and no NE for large s_R and s_P (see Figure 3a).

Let us now consider that the total amount of incentives is fixed at $C = R + P$. Proposition 3 points out that if $P + R/n \geq (n-r)E/n$, then the cooperative state is the only stable NE for larger s_P and s_R. This implies that $C \geq (n-r)E/n$ can sustain full cooperation. We next investigate the combination of reward and punishment that can maximize the group welfare. From Eq.(5), the group average payoff at the cooperative state is $Er + R/n$, i.e., reward can

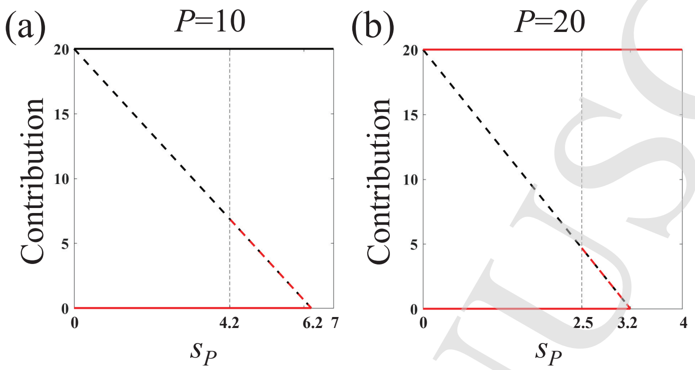

Figure 2: Institutional punishment. Parameters are taken as $E = 20$, $n = 4$, $r = 1.6$. NEs are denoted by red lines and stable equilibria of Eq.(6) are denoted by solid lines. **(a)** $P = 10$. The cooperative state is never a NE. The interior state and the free-riding state are no longer NEs for $s_P > 6.2$. **(b)** $P = 20$. The cooperative state is the only stable NE for $s_P > 3.2$.

help to increase the group welfare. Thus, the optimal amounts of punishment and reward are $P = [(n - r)E - C]/(n - 1)$ and $R = [nC - (n - r)E]/(n - 1)$, respectively (i.e., the maximum amount of R that can maintain full contribution), and the incentive institution should only punish the lowest contributor. Specifically, if $C \geq E(n - r)$, then reward alone is sufficient to maintain full contribution, and the use of punishment is redundant.

4 Conclusion

In this paper, we consider PGGs with three types of incentive institutions, namely reward, punishment, and a mixture of reward and punishment. For each type of incentive, we calculate NEs and analyse their evolutionary stability. The main purpose is to find the optimal incentive that is both effective in promoting cooperation and preserving group welfare. Overall, the effect of incentives on cooperation can be understood in terms of the size of the incentives. If the incentives are large, full contribution can be the unique NE for all the three incentives. However, if the incentives are of moderate size, the outcome depends crucially on the probability of being rewarded or punished. In the case of punishment, the institution should only punish the worst contributor (Yamagishi, 1986; Andreoni, Gee, 2012). Under the threat of punishment, rational players will try to avoid being the lowest contributor, so cooperation can be sustained. In contrast, reward can promote cooperation only if lower contributors also have a chance to be rewarded. In fact, if the institution only rewards the best contributor, then some subjects may give up on receiving the reward and free-ride. Finally, if the incentive institution can provide

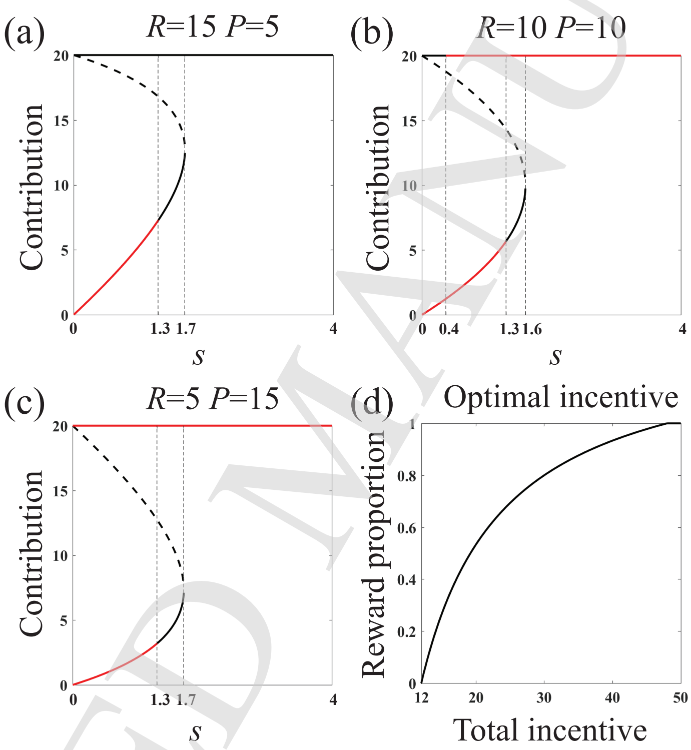

Figure 3: A mixture of reward and punishment. Parameters are taken as $E = 20$, $n = 4$, $r = 1.6$. Furthermore, we assume that $s_R = s_P = s$ in this figure. NEs are denoted by red curves, and stable equilibria of Eq.(6) are denoted by solid curves. **(a)** $R = 15$, $P = 5$. The cooperative state is never a NE. **(b)** $R = 10$, $P = 10$. The cooperative state is the only stable NE for $s > 1.3$. **(c)** $R = 5$, $P = 15$. The cooperative state is the only stable NE for $s > 1.3$. **(d)** The optimal combination of reward and punishment.

both reward and punishment, then it should use reward as much as possible. The group welfare is maximized when the punishment is just barely larger than the minimum required to obtain full contribution.

We now discuss some aspects of the theoretical models and review related literature.

In our model, the total amount of reward and/or punishment is fixed. It has been shown that adaptive incentives can also effectively promote cooperation, e.g., the penalty equals to the payoff difference between the lowest contributor and the second lowest contributor (Andreoni, Gee, 2012). In fact, our study provides a lower bound of the amount of incentives in sustaining cooperation, i.e., full contribution cannot be a NE if the total amount of reward and punishment is less than $(n-r)E/n$. Furthermore, we indicate that for any given amount of punishment, the best way to use it is to punish the worst contributor.

In addition, the incentives in our model are exogenous, i.e., both the reward and punishment are paid by the institution. Some recent studies considered the case of endogenous incentives. One class of studies assumed that subjects in the PGG have to pay a fee for the institution that will be used for reward and punishment (Sigmund et al., 2010; Sasaki et al., 2012; Traulsen et al., 2012; Chen et al., 2015; Yang et al., 2018; Dong et al., 2019). We note that this type of endogenous setting does not affect our results regarding the NEs and their stabilities, because adding a constant value to the payoff function will not change the adaptive dynamics. Moreover, the endogenous setting also does not qualitatively change the efficiencies of the different types of incentives (i.e., if IR is more efficient than IP in the exogenous setting, then it is also more efficient in the endogenous setting), although increasing the reward amount can no longer lead to a higher group payoff and increasing the punishment amount will decrease the social welfare. Another class of studies considered that subjects can voluntarily choose to join in the game and pay for reward or punishment (Kosfeld et al., 2009; Aimone et al., 2013; Zhang et al., 2014; Kopányi-Peuker et al., 2017; Lien, Zheng, 2019). If the payment is voluntary, the problem of second-order free-riding is raised because the incentive institution itself is a common good that can be exploited. The mechanism design in this case then becomes a more delicate issue.

Finally, our model assumes that the same amount of reward and punishment plays the same role in the payoff function. However, empirical studies observed that the responses to reward and punishment are often asymmetrically, where individuals are often more sensitive to losses than gains (Fehr, Goette, 2007; Dong et al., 2016; Lien et al., 2017). This phenomenon is captured by a type of reference-dependent preferences, called loss aversion (Kahneman, Tversky, 1979; Koszegi, Rabin, 2006; Knetsch et al., 2012; Eil, Lien, 2014; Lien, Zheng, 2015; Zhang, Zheng, 2017). Thus, a possible future development would be to incorporate loss aversion into the payoff function, and we expect that punishment can be more effective than reward.

In summary, our research deepens the understanding into the role of relative incentives in promoting cooperation. In particular, we show that the institution should use reward and punishment in different ways, i.e., punish the worst and do not only reward the best. This result is consistent with our life experience. In businesses, most of the employees who perform not too bad have the chance to get the bonuses (in many companies, bonuses is an important part of pay). Compared with reward, the use of punishment is less common. Most of employees do not face the risk of punishment, and only the worst employees will be penalised. Our results also suggest that the institution should use reward as much as possible, because reward is more efficient than punishment in promoting social welfare.

References

[1] Aimone, JA, LR Iannaccone, MD Makowsky and J Rubin (2013). Endogenous group formation via unproductive costs. *Rev. Econ. Studies*, 80, 1215-1236.

[2] Andreoni, J and LL Gee (2012). Gun for hire: delegated enforcement and peer punishment in public goods provision. *J. Public Econ.*, 96, 1036-1046.

[3] Chen, X, T Sasaki, Å Brännström and U Dieckmann (2015). First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation. *J R. Soc. Interface*, 12, 20140935.

[4] Chowdhury, SM and RM Sheremeta (2011). Multiple equilibria in Tullock contests. *Econ. Lett.*, 112, 216-219.

[5] Cressman, R, JW Song, B Zhang and Y Tao (2012). Cooperation and evolutionary dynamics in the public goods game with institutional incentives. *J. Theor. Biol.*, 299, 144-151.

[6] Cressman, R, JJ Wu, C Li and Y Tao Y (2013). Game experiments on cooperation through reward and punishment. *Biol. Theory*, 8, 158-166.

[7] Dieckmann, U and R Law (1996). The dynamical theory of coevolution: a derivation from stochastic ecological processes. *J Math. Biol.*, 34, 579-612.

[8] Dong, Y, C Li, Y Tao and B Zhang (2015). Evolution of conformity in social dilemmas. *PLoS ONE*, 10, e0137435.

[9] Dong, Y, T Sasaki and B Zhang (2019). The competitive advantage of institutional reward. *Proc. R. Soc. B*, 286, 20190001.

[10] Dong, Y, B Zhang and Y Tao (2016). The dynamics of human behavior in the public goods game with institutional incentives. *Sci. Rep.*, 6, 28809.

[11] Eil, D and JW Lien (2014). Staying ahead and getting even: Risk attitudes of experienced poker players. *Games Econ. Behav.*, 87, 50-69.

[12] Ewerhart, C (2015). Mixed equilibria in Tullock contests. *Econ. Theory*, 60, 59-71.

[13] Falkinger, J, E Fehr , S Gáchter and R Winter-Ebmer (2000). A simple mechanism for the efficient provision of public goods: Experimental evidence. *Am. Econ. Rev.*, 90, 247-264

[14] Fehr, E and L Goette L (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *Am. Econ. Rev.*, 97, 298-317.

[15] Galbiati, R and P Vertova (2008). Obligations and cooperative behaviour in public good games. *Games Econ. Behav.*, 64, 146-170.

[16] Hehenkamp, B, W Leininger and A Possajennikov (2004). Evolutionary equilibrium in Tullock contests: spite and overdissipation. *Eur. J. Political Econ.*, 20, 1045-1057.

[17] Hofbauer, J and K Sigmund (1998). *Evolutionary Games and Population Dynamics.* Cambridge University Press, Cambridge.

[18] Kahneman, D and A Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-292.

[19] Kamijo, Y, T Nihonsugi, A Takeuchi and Y Funaki (2014). Sustaining cooperation in social dilemmas: Comparison of centralized punishment institutions. *Games Econ. Behav.*, 84, 180-195.

[20] Knetsch, JL, YE Riyanto and Zong J (2012). Gain and loss domains and the choice of welfare measure of positive and negative changes. *J. Benefit Cost Anal.*, 3, 1-18.

[21] Kopányi-Peuker, A, T Offerman and R Sloof (2017). Fostering cooperation through the enhancement of own vulnerability. *Games Econ. Behav.*, 101, 273-290.

[22] Koszegi, B and M Rabin (2006). A model of reference-dependent preferences. *Q. J. Econ.*, 121, 1133-1165.

[23] Kosfeld, M, A Okada and A Riedl (2009). Institution formation in public goods games. *Am. Econ. Rev*, 99, 1335-1355.

[24] Lien, JW, M Xu and J Zheng (2017). What brings consumers back for more? Evidence from quantifiable gain and loss experiences in penny auctions. *Working Paper.*

[25] Lien, JW and J Zheng (2015). Deciding when to quit: Reference-dependence over slot machine outcomes. *Am. Econ. Rev*, 105, 366 - 370.

[26] Lien, JW and J Zheng (2019). Self-commitment for cooperation. *Working Paper.*

[27] Morgan, J (2000). Financing public goods by means of lotteries. *Rev. Econ. Studies*, 67, 761-784.

[28] Ostrom, E (2005). *Understanding institutional diversity.* Princeton: Princeton University Press.

[29] Putterman, L, JL Tyran and K Kamel (2011). Public goods and voting on formal sanction schemes. *J. Public Econ.*, 95, 1213-1222.

[30] Qin, X and S Wang (2013). Using an exogenous mechanism to examine efficient probabilistic punishment. *J. Econ. Psychol.*, 39, 1-10.

[31] Sigmund, K, H De Silva, A Traulsen A and C Hauert (2010). Social learning promotes institutions for governing the commons. *Nature* 466, 861-863.

[32] Sasaki, T, A Brännström, U Dieckmann U and K Sigmund (2012). The take-it-or-leave-it option allows small penalties to overcome social dilemmas. *Proc. Natl. Acad. Sci. USA.*, 109, 1165-1169.

[33] Traulsen, A, T Röhl and M Milinski (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B*, 279, 3716-3721.

[34] Tullock, G (1980). Efficient rent-seeking. In: Buchanan J, Tollison R, Tullock G (eds.) *Toward a Theory of the Rent-Seeking Society*, pp. 97-112. Texas A&M University Press, College Station.

[35] Wu, JJ, C Li, B Zhang B, R Cressman and Y Tao (2014). The role of institutional incentives and the exemplar in promoting cooperation. *Sci. Rep.*, 4, 6421.

[36] Yamagishi, T (1986). The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.*, 51, 110-116.

[37] Yang, C, B Zhang, G Charness, C Li and JW Lien (2018). Endogenous rewards promote cooperation. *Proc. Natl. Acad. Sci. USA*, 115, 9968-9973.

[38] Zhang, B, C Li, H De Silva, P Bednarik and K Sigmund (2014). The evolution of sanctioning institutions: an experimental approach to the social contract. *Exp. Econ.*, 17, 285-303.

[39] Zhang, M and J Zheng (2017). A robust reference-dependent model for speculative bubbles. *J. Econ. Behav. Organ.*, 137, 232-258.

Appendix 1: Proof for Proposition 1

Proof: For IR, Eq.(5) is written as

$$f_R(x, x^*) = E - x + \frac{r}{n}(x + (n-1)x^*) + \frac{Rx^{s_R}}{x^{s_R} + (n-1)x^{*s_R}}. \tag{7}$$

The cooperative state $x^* = E$ is a NE if and only if $f_R(x, E) \leq f_R(E, E)$ for all $0 \leq x < E$. Thus, two necessary conditions for the cooperative NE are (i) $f_R(0, E) \leq f_R(E, E)$ and (ii) $\frac{\partial f_R(x,E)}{\partial x}|_{x=E} \geq 0$. Condition (i) implies $R \geq (N - r)E$, and condition (ii) is equivalent to

$$\frac{\partial f_R(x, E)}{\partial x}|_{x=E} = -1 + \frac{r}{n} + \frac{Rs_R(n-1)}{n^2 E} \geq 0, \tag{8}$$

where the inequality holds for $s_R \geq \frac{n}{n-1}$ (with condition (i) $R \geq (N - r)E$). In addition, if $s_R \geq \frac{n}{n-1}$ (i.e., condition (ii) holds), then

$$\frac{\partial^2 f_R(x, E)}{\partial x^2} = \frac{Rs_R(n-1)E^{S_R}((s_R - 1)(n-1)E^{s_R} - x^{s_R})}{(x^{s_R} + (n-1)E^{s_R})^3} \geq 0 \tag{9}$$

for all $x \in [0, E]$. This implies that $f_R(x, E)$ takes it maximum at either $x = 0$ or $x = E$, i.e., $x^* = E$ is a NE if condition (i) holds. Thus, conditions (i) and (ii) are also sufficient for the cooperative NE.

On the other hand, the cooperative state is never a NE for $R < (n-r)E$, and an interior NE may exist. An interior state x^* is a NE only if it is a local maximum point of $f_R(x, x^*)$. Thus,

$$\begin{aligned}
\frac{\partial f_R(x, x^*)}{\partial x}|_{x=x^*} &= -1 + \frac{r}{n} + \frac{Rs_R(n-1)}{n^2 x^*} = 0, \\
\frac{\partial^2 f_R(x, x^*)}{\partial x^2}|_{x=x^*} &= \frac{Rs_R(n-1)(ns_R - 2s_R - n)}{n^3 x^{*2}} \leq 0,
\end{aligned}$$

where the first equality yields $x^* = \frac{Rs_R(n-1)}{n(n-r)}$ and the second inequality implies $s_R \leq \frac{n}{n-2}$. Furthermore, when all other subjects are using strategy x^*, a subject deviating to free-riding should not obtain a higher payoff, i.e., $f_R(0, x^*) \leq f_R(x^*, x^*)$. This yields $x^* = \frac{Rs_R(n-1)}{n(n-r)} \leq \frac{R}{n-r}$, or equivalently $s_R \leq \frac{n}{n-1}$. Overall, the interior state $x^* = \frac{Rs_R(n-1)}{n(n-r)}$ is a NE only if $s_R \leq \frac{n}{n-1}$.

For the evolutionary stability, if $R < \frac{n(n-r)E}{s_R(n-1)}$, Eq.(6) has a unique interior fixed point $x^* = \frac{Rs_R(n-1)}{n(n-r)}$, which coincides with the interior NE. Furthermore, it is globally stable since $\frac{dx}{dt} > 0$ for $x < x^*$ and $\frac{dx}{dt} < 0$ for $x > x^*$. On the other hand, if $R \geq \frac{n(n-r)E}{s_R(n-1)}$, then $\frac{dx}{dt} \geq 0$ for all $0 < x \leq E$ and the cooperative state $x^* = E$ is globally stable. This implies that a NE in IR must be globally stable under adaptive dynamics. □

Appendix 2: Proof for Proposition 2

Proof: For IP, Eq.(5) is written as

$$f_P(x, x^*) = E - x + \frac{r}{n}(x + (n-1)x^*) - \frac{P(E-x)^{s_P}}{(E-x)^{s_P} + (n-1)(E-x^*)^{s_P}} \tag{10}$$

The cooperative state $x^* = E$ is a NE if and only if decreasing the contribution cannot obtain a higher payoff, i.e., $f_P(x, E) \leq f_P(E, E)$ for all $x < E$. Since we always have $f_P(0, E) \geq f_P(x, E)$, $x^* = E$ is a NE if and only if $f_P(0, E) \leq f_P(E, E)$. Thus, we obtain $P \geq \frac{E(n-r)}{n}$.

The free-riding state $x^* = 0$ is a NE only if a subject deviates to cooperation or slight increases in the contribution cannot obtain higher payoff, i.e., $f_P(E, 0) \leq f_P(0, 0)$ and $\frac{\partial f_P(x,0)}{\partial x}|_{x=0} \leq 0$. The first inequality implies $P \leq (n-r)E$, and the second inequality implies $P \leq \frac{nE(n-r)}{s_P(n-1)}$.

Finally, an interior state x^* is a NE only if it is a local maximum point of $f_P(x, x^*)$. Thus,

$$\frac{\partial f_P(x, x^*)}{\partial x}|_{x=x^*} = -1 + \frac{r}{n} + \frac{P s_P(n-1)}{n^2(E - x^*)} = 0,$$
$$\frac{\partial^2 f_P(x, x^*)}{\partial x^2}|_{x=x^*} = -\frac{P s_P(n-1)(n s_P - 2 s_P - n)}{n^3(E - x^*)^2} \leq 0,$$

where the first equality yields $x^* = E - \frac{P s_P(n-1)}{n(n-r)}$ and the second inequality implies $s_P \geq \frac{n}{n-2}$. From the boundary condition $x^* > 0$, we obtain $s_P < \frac{nE(n-r)}{P(n-1)}$. Finally, when all other subjects are using strategy x^*, deviating to full contribution cannot obtain a higher payoff, i.e., $f_P(E, x^*) \leq f_P(x^*, x^*)$. This yields $x^* = E - \frac{P s_P(n-1)}{n(n-r)} \leq E - \frac{P}{n-r}$, or equivalently, $s_P \geq \frac{n}{n-1}$. Overall, a sufficient condition for the interior NE is $\frac{n}{n-2} \leq s_P < \frac{nE(n-r)}{P(n-1)}$. In addition, if $s_P \geq \frac{nE(n-r)}{P(n-1)}$, then $x^* = E$ is the unique NE.

For the evolutionary stability, if $P \geq \frac{n(n-r)E}{s_P(n-1)}$, the cooperative state $x = E$ becomes the only stable fixed point, and it is also globally stable. If $0 < P < \frac{n(n-r)E}{s_P(n-1)}$, Eq.(6) has a unique interior fixed point $x^* = E - \frac{P s_P(n-1)}{n(n-r)}$, which coincides with the interior NE. Furthermore, $\frac{dx}{dt} < 0$ for $x < x^*$ and $\frac{dx}{dt} > 0$ for $x > x^*$. This implies that the interior fixed point must be unstable. In this case, both the free-riding state $x = 0$ and the cooperative state $x = E$ are locally stable fixed points. Specifically, increasing s_P can increase the basin of attraction of the cooperative fixed point. As s_P goes to infinity, the interior and the free-riding fixed points vanish, and the cooperative fixed point becomes globally stable. □

Appendix 3: Proof for Proposition 3

Proof: The cooperative state $x^* = E$ is a NE if and only if $f_{RP}(x, E) \leq f_{RP}(E, E)$ for all $x \in [0, E]$, i.e.,

$$f(x, E) - f(E, E) = E - x + \frac{r}{n}(x - E) + \frac{Rx^{s_R}}{x^{s_R} + (n-1)E^{s_R}} - \frac{R}{n} - P \leq 0. \tag{11}$$

Eq.(11) is independent of s_P. It can be simplified as $\frac{E(n-r)}{n} - P - \frac{R}{n} \leq 0$ as $s_R \to \infty$, which means that $x^* = E$ is a NE for $P + \frac{R}{n} \geq \frac{E(n-r)}{n}$. On the other hand, $f_{RP}(x, E) - f_{RP}(E, E) \leq E - x + \frac{r}{n}(x - E) - \frac{R}{n} - P$, which means that $x^* = E$ is a NE for $P \geq \frac{E(n-r)}{n}$.

An interior state x^* is a NE only if it is a local maximum point of $f_R P(x, x^*)$. Thus,

$$\begin{aligned}
\frac{\partial f(x, x^*)}{\partial x}|_{x=x^*} &= -1 + \frac{r}{n} + \frac{Rs_R(n-1)}{n^2 x^*} + \frac{Ps_P(n-1)}{n^2(E - x^*)} = 0, \\
\frac{\partial^2 f(x, x^*)}{\partial x^2}|_{x=x^*} &= (n-1)\left[\frac{Rs_R(ns_R - 2s_R - n)}{n^3 x^{*2}} + \frac{Ps_P(2s_P - ns_P + n)}{n^3(E - x^*)^2}\right] \leq 0.
\end{aligned}$$

Notice that $\frac{\partial f(x,x^*)}{\partial x}|_{x=x^*} = 0$ is a second-order equation in x^*, it can have at most two solutions. However, when $s_R \geq \frac{n(n-r)E}{R(n-1)}$ or $s_P \geq \frac{n(n-r)E}{P(n-1)}$, it has no solution and $\frac{\partial f(x,x^*)}{\partial x}|_{x=x^*} > 0$ for all $x^* \in [0, E]$.

For the evolutionary stability, Eq.(6) can have at most two interior fixed points. We denote them by $0 < x_1^* < x_2^* < E$. Notice that $\frac{dx}{dt} > 0$ for $x \to 0$ and $x \to E$, $x = 0$ is not a fixed point and $x = E$ is a stable fixed point. From the continuity of $\frac{dx}{dt}$, we must have $\frac{\partial dx/dt}{\partial x} < 0$ for $x = x_1^*$ and $\frac{\partial dx/dt}{\partial x} > 0$ for $x = x_2^*$. This implies that $x = x_1^*$ is locally stable, $x = x_2^*$ is unstable, and $x = E$ is locally stable. Furthermore, if Eq.(6) does not have interior fixed point, then $\frac{dx}{dt} > 0$ for all $x \in [0, E]$, which means that the cooperative fixed point is globally stable. □