Game Experiments on Cooperation Through Reward and Punishment

Ross Cressman, Jia-Jia Wu, Cong Li & Yi Tao

Biological Theory

ISSN 1555-5542

Biol Theory
DOI 10.1007/s13752-013-0106-2

Biol Theory
DOI 10.1007/s13752-013-0106-2

THEMATIC ISSUE ARTICLE: STRATEGIC INTERACTION

Game Experiments on Cooperation Through Reward and Punishment

Ross Cressman · Jia-Jia Wu · Cong Li · Yi Tao

Received: 1 December 2011 / Accepted: 9 October 2012

Abstract Game experiments designed to test the effectiveness of reward and/or punishment incentives in promoting cooperative behavior among their participants are quite common. Results from two such recent experiments conducted in Beijing, based on the Prisoner's Dilemma (PD) game and Public Goods Game respectively, are summarized here. The unexpected empirical outcomes for the repeated PD game, that cooperation actually decreased when the participants had the option of using a costly punishment strategy and that participants who used costly punishment in some round of the repeated game often did so in the first round, are discussed in terms of differences in attitudes toward reputation in Chinese culture compared to other locations (mostly in Western society) where similar experiments have been conducted. The second experiment models an institution providing incentives to increase contribution levels (i.e., cooperation) to the public good. The results show that combined institutional reward and punishment is the most effective means to increase cooperation, followed by a scheme using only punishment. It is shown how these empirical results are related to the theoretical predictions that assume players play rationally by optimizing their personal payoff given their opponents' actions in these multi-player games.

R. Cressman (✉)
Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: rcressman@wlu.ca

J.-J. Wu · C. Li · Y. Tao
Key Lab of Animal Ecology and Conservational Biology, Centre for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China
e-mail: yitao@ioz.ac.cn

Keywords Costly punishment · Cultural effects · Institutional incentives · Prisoner's Dilemma · Public Goods Game

Introduction

A great deal of theoretical (e.g., Sigmund 2010; Sigmund et al. 2001) and experimental (e.g., Andreoni et al. 2003) research has examined the effectiveness of punishment and/or reward incentives in promoting cooperative behavior in humans. These studies often take the underlying one-shot base game as either the two-player Prisoner's Dilemma (PD) with strategies Cooperate (*C*) and Defect (*D*) or the multi-player Public Goods Game (PGG) where each player chooses how much of his endowment to contribute to a common pool. In both base games, it is individually advantageous to defect no matter what others do (i.e., Defect is the only Nash equilibrium [NE] of the game) whereas the best collective outcome occurs when everyone cooperates (i.e., Cooperate is the socially optimal strategy).[1]

Experiments on PD and PGG do not agree with the NE prediction that assumes individually rational behavior (Ledyard 1995). Instead, observed cooperative behavior can be over 50 % initially (i.e., average contribution more than half the endowment in PGG and more than half the participants play C in PD) and typically tends to decrease to some lower bound above 20 % as individuals play the game more often. Interestingly, these ranges for cooperative behavior do not seem to depend on whether the games are repeated with the same players or opponents in each game are different.

[1] In PGG, Cooperate corresponds to contributing the entire endowment and Defect to contributing nothing.

Published online: 25 April 2013

Given these results, we do not expect NE behavior to emerge in experiments that add incentives to cooperate to the base game. On the other hand, there is evidence that observed behavior can be described through a combination of maximizing one's own income given the behavior of others (i.e., through NE behavior) and other intrinsic motivations such as a sense of caring about the well-being of others or working together for the common good (Schram 2000). What is then of interest is how these incentives change the NE prediction and to what extent this is reflected in observed levels of cooperation.

This paper is a synthesis of two game experiments conducted in Beijing on PD and on PGG (Tao et al. 2009; Wu et al. 2009, 2012) together with a theoretical analysis (Cressman et al. 2012) that connects evolutionary dynamics to the NE outcome of PGG augmented by institutional incentives. "Costly Punishment and the PD Game" section summarizes the repeated PD game experiment (Wu et al. 2009) with and without the option of costly punishment. Two unexpected outcomes are reported there and discussed in terms of the participants' attitude to reputation. "Institutional Incentives and PGG" section introduces the repeated PGG experiment (Wu et al. 2012) with and without institutional incentives. It begins with the NE analysis of the one-shot games (Cressman et al. 2012) and then connects these results with the experimental outcomes reported in "Game Experiment Based on PGG" section. The Conclusion ("Conclusion" section) provides a brief discussion of the main highlights of these experiments and their game-theoretic properties.

Costly Punishment and the PD Game

Our first experimental study (Wu et al. 2009) is based on the two-player game with options A and B whose payoff consequences are

Option	You will get	The other person will get
A	$-c$	b
B	d	$-d$

for the positive parameters b, c, d. With option A and B corresponding to Cooperate (C) and Defect (D) respectively, the payoff matrix becomes

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} b-c & -c-d \\ b+d & 0 \end{pmatrix} \end{array}. \qquad (1)$$

This game can then be reinterpreted as the standard PD game where a Cooperator incurs a cost $c+d$ to himself and gives a benefit $b+d$ to the other player. Clearly, the only NE is to Defect since D strictly dominates C.

However, given that some players do Cooperate in lab experiments using these games, one would expect more cooperation in games with higher b (when c, d are unchanged) if players use other intrinsic motivations to choose their strategy. Specifically, the higher b is, the more a Cooperator increases the well-being of the other player and also the common good (since both entries in the first column of the payoff matrix (1) increase).

In the two treatments of our experiment, a third option is added whereby the player gets $-\alpha$ for the other person to get $-\beta$. This corresponds to costly punishment (P) (also called altruistic punishment) since the player pays the cost α for the other player to incur a penalty β.[2] The payoff matrix is now

$$\begin{array}{c} \\ C \\ D \\ P \end{array} \begin{array}{c} \begin{array}{ccc} C & D & P \end{array} \\ \begin{pmatrix} b-c & -c-d & -c-\beta \\ b+d & 0 & d-\beta \\ b-\alpha & -\alpha-d & -\alpha-\beta \end{pmatrix} \end{array}. \qquad (2)$$

Again, D strictly dominates the other two strategies and so the only NE is Defect.

It is often argued that costly punishment is an effective peer-incentive method to promote cooperation among humans (Baldassarri and Grossman 2011), especially when the cost to the punisher is (substantially) less than the penalty to the punished (i.e. $\alpha < \beta$). For instance, if the game is repeated between the same two people, a player who defects in one round can be subsequently punished to encourage him to change his strategy. This type of prosocial punishment has been observed in many experiments (e.g., Herrmann et al. 2008), resulting in increased levels of cooperation. However, this is not the case in our experiment where players knew that, at the end of each round, there would be another round with probability 0.75. The two controls based on (1) have parameters $c = 1 = d$ and either $b = 2$ (C1) or $b = 3$ (C2). The two treatments, T1 and T2, based on (2) use these same parameters together with $\alpha = 1$ and $\beta = 4$.

From Fig. 1, there is no significant difference in the level of cooperation among these four sessions (see Wu et al. 2009 for the statistical analysis), with mean levels ranging from 18.2 to 26.4 %. In particular, the average frequency of cooperation does not increase in the control sessions with higher b (C1 vs. C2). Furthermore, the average frequency of cooperation actually decreases with the addition of costly punishment when $b = 2$ (C1 vs. T1). Both of these results are counter to what is expected from the above discussion. Moreover, they are opposite to those obtained in the identical experiment conducted in Boston by Dreber et al. (2008) where the average frequencies of cooperation were 21.2 % (C1), 52,4 % (T1), 43.0 % (C2) and 59.7 % (T2).

[2] From this perspective, option A (i.e., Cooperate) could be called costly reward instead since the player pays a cost c for the other player to receive the benefit b.

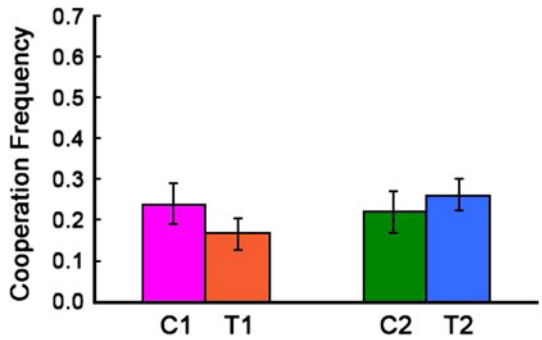

Fig. 1 Average frequency of cooperation in each session of the repeated PD game experiment (Wu et al. 2009), without costly punishment, C1 ($b = 2$) and C2 ($b = 3$), and with costly punishment, T1 ($b = 2$) and T2 ($b = 3$). *Error bars* represent the standard error from the mean. Other parameters are $c = 1 = d$, $\alpha = 1$ and $\beta = 4$

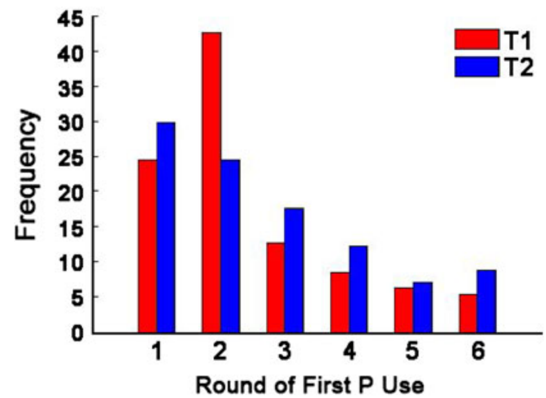

Fig. 2 The round when P is used first in those treatments where it is used sometime

Two possible explanations for our unexpected results are advanced by Wu et al. (2009). The first rests on the finding (e.g., Standifird 2006) that good reputations are built in Chinese culture through multi-player (guanxi) networks rather than through two-person (dyadic) interaction (which are more important in Western society). It is then argued that participants in our two-player experiment in Beijing are less concerned about adjusting their baseline cooperative behavior due to other intrinsic motivations than in Boston, where pro-social punishment may be more directed at players with a bad reputation.

This latter conclusion is supported by the second explanation that considers when players who punish first use this strategy. Of the approximately 6 % of participants who use P sometime in our repeated PD game experimental treatments, 24.4 and 29.8 % use it in the first round of T1 and T2 respectively (see Fig. 2). That is, there is substantial first use of P before observing any behavior of the other player. These participants are punishing indiscriminately in an unprovoked first strike, perhaps in an attempt to establish their reputation as a dominant authority figure (Wu et al. 2009). This result contrasts markedly with the experiment in Boston where, in both T1 and T2, P is used in the first round in less than 5 % of the cases where it is used sometime in the treatment (Dreber et al. 2008). This discrepancy between the two locations again points to cultural differences in attitudes toward reputation.

On the other hand, in both locations, the option of costly punishment does not increase the average payoff from that obtained by the participants in the control sessions.[3] This outcome is typical of experiments involving peer punishment (Fehr and Gächter 2000) in that any payoff gain in cases where cooperation increases is offset by decreased payoff from punishment. It is then speculated that costly punishment has evolved in human societies (and elsewhere) for other reasons than to promote cooperation (Dreber et al. 2008), a conclusion that remains controversial (Rankin et al. 2009; Tao et al. 2009). One of the alternative reasons given is that costly punishment may have evolved as a means of "coercing individuals into submission and establishing dominance hierarchies" (Sherratt and Wilkinson 2009), thereby sustaining cooperation once this behavior is established (Mathew and Boyd 2011).

Institutional Incentives and PGG

In our second experiment (Wu et al. 2012), it is assumed that the incentives to cooperate are already established in the society as a formal institutionalized scheme that either rewards desirable behavior or punishes undesirable behavior (or both). Moreover, in contrast to the peer-incentive punishment scheme discussed in the "Costly Punishment and the PD Game" section, it is the institution that bears any cost in administering the incentive.

For example, many firms regularly rank the performance of their employees, rewarding those who perform well and sanctioning (or punishing) poorly performing individuals (Prendergast and Topel 1993). An extreme implementation of such a scheme is a forced ranking system (Grote 2005) whereby employees ranked in the bottom 10 % are terminated and those in the top 20 % are given monetary bonuses or other rewards. As another example, governmental institutions enact many types of incentive schemes, ranging from punishment levels for different types of criminal behavior set by the judicial system to prestigious awards given to exemplary citizens.

Our game-theoretic model is based on the n-player PGG with $n \geq 2$. In this single-round game, each player in the group is given a fixed endowment $E > 0$ and contributes as much of this $x \in [0, E]$ as she wants to a common pool. The total amount in the pool is multiplied by a factor $r > 1$ and then redistributed evenly to each player in the group. We

[3] In our T1, there is in fact a highly significant decrease in average payoff compared with C1 (Wu et al. 2009).

assume that the player receives only a fraction of her own contribution to the common pool (i.e. $\frac{r}{n} < 1$). That is, the player's return on his contribution is less than 100 %. It is to the group's advantage if all players Cooperate (i.e., contribute E) but each player, given the contributions of the others, does best to Defect (i.e., free rides by contributing nothing). In particular, the only NE of PGG is for all players to free ride.

PGG is an example of an n-player game that has the form of a population game since a player's payoff $\pi(x; x_2, x_3, \ldots, x_n) = E + \left(\frac{r}{n} - 1\right)x + \frac{r}{n}\sum_{i=2}^{n} x_i$ depends only on his strategy x and the average strategy $x_{-1} \equiv \frac{1}{n-1}\sum_{i=2}^{n} x_i$ of the rest of his group. Moreover, it is a symmetric game (in the sense that payoffs do not depend on designating who is player 1, who is player 2, etc.) with a continuum of pure strategies given by the one-dimensional interval $[0, E]$.[4] The continuous strategy spaces of PGG, with or without incentives, makes their game-theoretic analysis more interesting. Cressman et al. (2012) investigated the standard evolutionary dynamics for these games; namely, the canonical equation of adaptive dynamics (Dieckmann and Law 1996) and the replicator equation for a continuous strategy space (Cressman and Hofbauer 2005). When there are no incentives, the evolutionary outcome is the free-riding NE (i.e., $x = 0$ is the globally asymptotically stable equilibrium of both dynamics).

Institutional incentives are included in the model as follows. After the standard PGG is played among the group of n players, a second stage is added where one group member is chosen to be rewarded (institutional reward, IR) or one group member is chosen to be punished (institutional punishment, IP) or both (institutional reward and punishment, IRP). The incentive amount (i.e., the additional payoff given to the member in IR or taken away in IP) is fixed at $A > 0$.[5] The probability of being chosen for a reward and/or punishment depends on the member's contribution and those of the rest of the group.

The following three results for IR, IP, and IRP respectively summarize the NE analysis in Cressman et al. (2012) for such models that include the game experiment based on PGG of the "Game Experiment Based on PGG" section.

(a) Institutional reward. In IR, assume that exactly one individual receives the reward and that this probability F depends only on her contribution x and the average contribution x_{-1} of the rest of the group. Specifically, take $F(x; x_{-1}) \equiv \frac{x+1}{x+(n-1)x_{-1}+n}$. Then, the incentive encourages each individual to be more cooperative in order to have a better chance of getting the reward (i.e., F is an increasing function of x) and, if all members contribute the same amount, each is equally likely to receive the reward (i.e., $F = \frac{1}{n}$). For instance, if $n = 4$ and $E = 20$ (which, along with $r = 1.6$, are commonly used parameters in experiments involving PGG (e.g., Herrmann et al. 2008; Rand et al. 2009), a free rider in a group that otherwise cooperates has only 1 chance in 64 of getting the reward versus 1 chance in 4 if to she switches to Cooperate.
For these parameters, Cressman et al. (2012) show that there is only one NE in which everyone contributes the same amount x^* (i.e. $\pi(x; x^*) \leq \pi(x^*; x^*)$ for all $x \in [0, E]$ where $\pi(x; x^*) = E + \left(\frac{r}{n} - 1\right)x + \frac{r(n-1)}{n}x^* + \frac{x+1}{x+(n-1)x_{-1}+n}A$). This depends on the reward amount A according to

$$x^* = \begin{cases} 0 & \text{if } A \leq 3.2 \\ \frac{5A}{16} - 1 & \text{if } 3.2 < A < 67.2 \\ 20 & \text{if } A \geq 67.2 \end{cases} \tag{3}$$

That is, rewards are not sufficient to offset the advantage of free riding unless the amount is at least 3.2, after which an intermediate NE level of cooperation emerges that linearly increases in A until finally Cooperate is the only NE once $A = 67.2$.

(b) Institutional punishment. In IP, assume that the probability of punishment is symmetric about the interval $[0, E]$ with the probability of reward in IR. That is, the probability $G(x; x_{-1})$ that an individual who contributes x is punished in a group whose other members contribute x_{-1} on average is the same probability that an individual who contributes $E - x$ is rewarded in IR when the other members of his group contribute $E - x_{-1}$ on average. From Cressman et al. (2012) with parameters $n = 4$, $E = 20$ and $r = 1.6$, free riding is a NE if $A \leq 67.2$. However, Cooperate becomes a NE as soon as $A \geq 3.2$ and no intermediate level of cooperation is ever a NE. In particular, for all punishment amounts between 3.2 and 67.2, both Cooperate and Defect are the NE.

(c) Institutional reward and punishment. In IRP, one group member is chosen to be rewarded with probability F and, independently, one to be punished[6] with probability G. For parameters $n = 4$, $E = 20$ and $r = 1.6$, Defect (Cooperate) is the only NE when the value of A is small (large). For values in between, there are two NE (corresponding to Cooperate and an intermediate NE level of cooperation).

[4] The two-player games of the "Costly Punishment and the PD Game" section with payoff matrices (1) and (2) are also symmetric but have only finitely many pure strategies (two and three respectively).

[5] In our second experiment (Wu et al. 2012), there are actually three protocols, one of which (Const) takes A fixed. The two other are Up (Down) where A increases (decreases) with higher group contributions.

[6] In particular, this could be the same person.

When $A = 20$ as in the game experiment, these three NE results summarized above are part of Table 1. For instance, we see that this incentive is large enough to imply Cooperate is the only NE in IRP. The table also includes, for $A = 20$, the evolutionary outcome under the canonical equation of adaptive dynamics that assumes all individuals use the same strategy $x \in [0, E]$ (i.e., the resident population is monomorphic) and that this monomorphism evolves through trait substitution in the direction of nearby strategies that have higher payoff than the resident strategy when played against the resident population. When there is a single NE (in IR and IRP), we see that it is globally asymptotically stable. On the other hand, when there are two NE (in IRP), there is another unstable rest point $\widehat{x}$ of the canonical equation between them. In fact, this rest point determines the domains of attraction of the NE. If the initial level of cooperation is above (below) $\widehat{x}$, the evolutionary dynamics evolves to Cooperate (Defect).

Evolution under the replicator equation is more complicated since the population is no longer assumed to be monomorphic. Instead, the distribution of strategies evolves through payoff comparisons by increasing the relative frequency of the strategy with higher payoff. However, numerical simulations (Cressman et al. 2012) agree qualitatively with the canonical equation. In particular, in IR and IRP, the limiting distribution (i.e., the evolutionary outcome) has everyone playing the single NE. For IP, the limit is again one of the NE but which one it is depends on the initial distribution (see Fig. 3).

Table 1 shows clearly that a positive incentive (reward) can have quite a different effect on NE behavior than a negative incentive (punishment) (cf. Rand et al. 2009). The predictions are:

(i) IR is able to increase cooperation when contribution levels are low but cannot maintain cooperation when levels are high;
(ii) IP cannot increase cooperation when levels are low but maintains it when high;
(iii) IRP combines both features, increases contributions when they are low and maintains them when high so that Cooperate is the only NE.

Table 1 NE, rest points and dynamic stability under the canonical equation when $A = 20$

Incentive schemes	Rest point	Nash equilibrium	Dynamic stability
IR	5.25	Yes	Globally stable
IP	0	Yes	Locally stable
	14.75	No	Unstable
	20	Yes	Locally stable
IRP	20	Yes	Globally stable

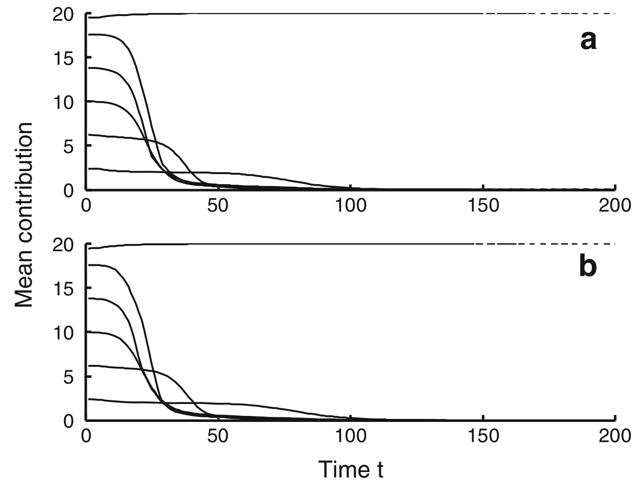

Fig. 3 The replicator equation (panel **a**) and its mean-field approximation (panel **b**) for IP with parameters as in Table 1. These panels show the mean contributions along six trajectories whose initial distributions have support on the strategies $x_i = i$ for $i = 0, 1, \ldots, 20$ corresponding to integer contributions. These initial distributions place 95 % of their weight on one strategy ($x = 20, 18, 14, 10, 6, 2$ respectively) and the remaining 5 % uniformly on the other 20 integer strategies. The mean field approximates the replicator equation by comparing the expected payoff of an individual using strategy x_i to the average payoff of the population instead of the average payoff of the randomly chosen group of n-players in which he is a member. The results in **a** and **b** are qualitatively very similar (in particular, both have the same two limiting distributions) but not completely identical (Cressman et al. 2012)

Game Experiment Based on PGG

In our second game experiment, groups of four play the repeated PGG (some with and some without incentives) for 50 rounds, knowing the contributions made and the payoffs received by each player in the previous round.[7] Parameters are $r = 1.6$, $E = 20$ and, when there are incentives, reward and punishment probabilities, F and G, are given in the above theoretical summary taken from Cressman et al. (2012). The protocol Const has fixed incentive $A = 20$ as in the above theoretical summary, Up has the incentive increasing linearly in total group contribution from 16 (when all Defect) to 25.6 (when all Cooperate) and Down has the incentive decreasing linearly in total group contribution from 25.6 (when all Defect) to 16 (when all Cooperate).

We will focus our discussion on how the three treatments in Const affect cooperation compared to PGG without incentives (i.e., the Control). As expected, average contribution levels in the experiment (Fig. 4) do not match the NE prediction of Table 1. Overall, IRP is clearly the most effective at promoting cooperation, followed by IP, and finally IR (which is better than the Control (Ctrl) on

[7] See Wu et al. (2012) for the detailed experimental design. For instance, the participants were not told that there were exactly 50 rounds.

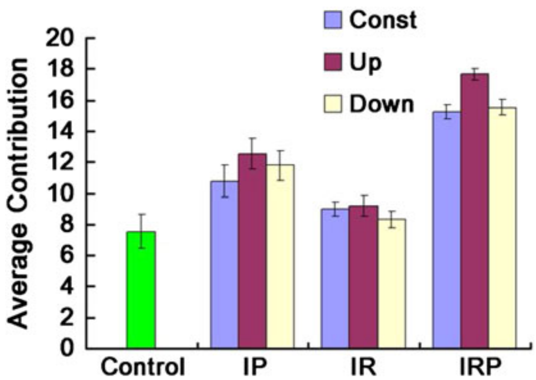

Fig. 4 Average contribution for the 50 round sessions in the Control and the 9 treatments. These are 7.55 in Ctrl and, for Const, 10.74 in IP, 8.95 in IR, 15.53 in IRP. *Error bars* represent the standard error from the mean

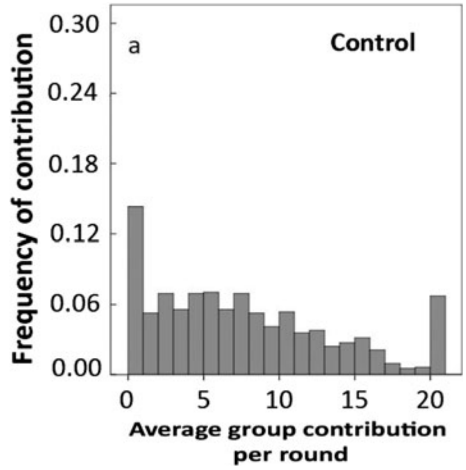

Fig. 5 Frequency distribution of average group contribution per round in the Control

average but not significantly so). The frequency distribution (Fig. 5) in Ctrl of the average group contribution per round is typical of game experiments on PGG (e.g., Rand et al. 2009) with a few groups contributing (close to) the collective optimal of 20 but many more groups near the free-riding NE. What is interesting is to see how this compares to the distribution in the treatments.

Empirically, there are fewer groups with low average contribution levels in all nine incentive treatments (Fig. 6) compared to Ctrl. That is, overall, the incentives are promoting cooperative behavior in groups. In IR, the distribution of group contributions is concentrated more in the mid-range level (consistent with point i above in relation to Table 1); whereas the trend in IRP is clearly to high levels (point iii). In IP, there are more groups at the high end compared to IR and C (points i and ii) but the distribution also maintains many groups at the low end compared to IRP (points ii and iii).

When there is a single NE (i.e., in Ctrl, IR, or IRP), the results agree qualitatively with other experiments on PGG that include mechanisms to move the NE away from Defect (Falkinger et al. 2000; Isaac and Walker 1998). In particular, observed contributions are above NE behavior when the NE is low (in Ctrl and IR) and below NE behavior when the NE is high (in IRP). The experimental results become understandable by combining Table 1 with intrinsic motivations that move behavior away from these extreme NE through factors other than pure self-interest.

This insight also emerges by considering what an individual should do in one round of our experiments if all other members of his group Cooperate (Fig. 7). In Ctrl, his expected relative payoff[8] to that of the other group members decreases from 12 to 0 as his relative contribution increases from −20 (Defect) to 0 (Cooperate). Since the same decreasing pattern (i.e., the higher relative contribution an individual makes, the smaller his expected relative payoff) holds in IR, the rational self-interested individual should free ride. Here, the 25 % chance of receiving the reward amount by cooperating is outweighed by the savings from free riding. However, in IP and IRP, the incentive to cooperate is much higher since he then has only a 25 % chance of being punished versus near certainty if he free rides. In fact, his expected relative payoff is now maximized by contributing 20 and so Cooperate is a NE. The intuition from these theoretical results is reflected in the experimentally observed average contribution levels in both IP and IRP being higher than in either IR or Ctrl. These theoretical calculations from Wu et al. (2012) also illustrate a major difference with typical peer-incentive PGG experiments (Fehr and Gächter 2000; Herrmann et al. 2008) where an individual has equivalent incentives to switch from Defect to Cooperate in the punishment treatment and in the reward treatment; namely, attempting to avoid punishment from each of the other group members versus gaining the reward from each of them. Institutional incentive schemes such as ours avoid the anti-social punishment that often emerges in peer-incentive schemes (Herrmann et al. 2008).[9]

The theoretical analysis summarized in the preceding paragraph can be used to give a particularly revealing

[8] Expected relative payoff is the difference between his expected payoff given his current contribution and given that he changes his contribution to the average of the other group members.

[9] Anti-social punishment refers to peers punishing high contributors (e.g., Cooperators) as opposed to pro-social punishment mentioned in the "Costly Punishment and the PD Game" section where peers punish Defectors. The levels of anti-social punishment in PGG with peer incentives are culturally dependent, although they appear to be about the same in China as in Western society (Herrmann et al. 2008). The first round pre-emptive punishment reported in Wu et al. (2009) for the PD game with costly punishment is neither anti- nor pro-social in the strict sense, although it is heursitically closer to the former.

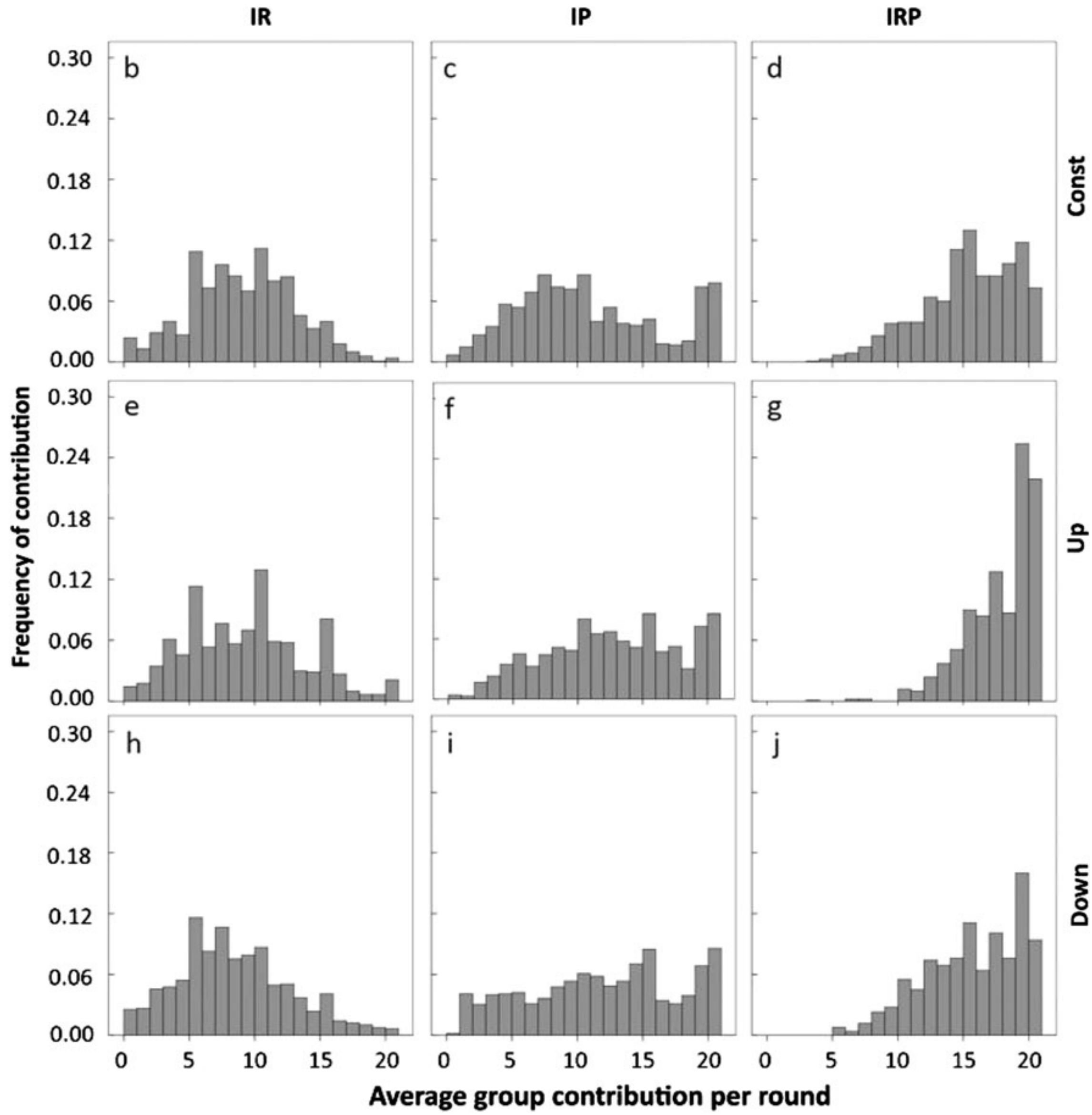

Fig. 6 Frequency distribution of average group contributions per round in IR, IP and IRP

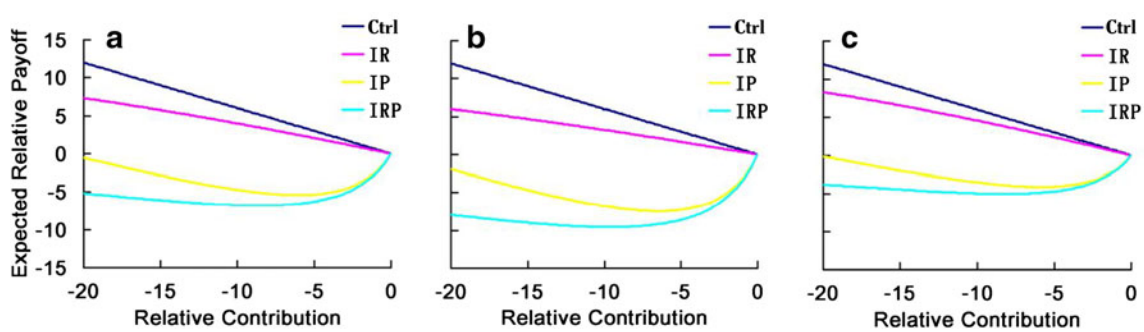

Fig. 7 Expected relative payoff in a group where all other members Cooperate in Const (**a**), Up (**b**) and Down (**c**). The individual's relative contribution increases from −20 when he Defects to 0 when he Cooperates

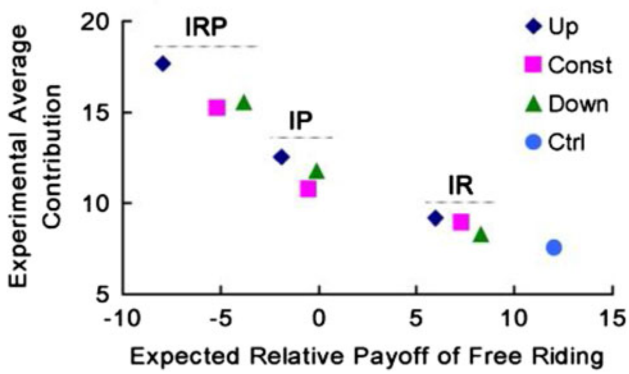

Fig. 8 Observed average contribution versus the expected relative payoff of free riding in a group whose other members Cooperate

empirical relationship between observed contribution level and the expected relative payoff of free riding when all other group members Cooperate. From Fig. 8, as the material benefits of free riding increase, there is a clear tendency to decrease contributions. This result supports the conclusion that participants are responding directly to appropriately sized monetary incentives.

Conclusion

Game experiments involving reward and punishment have become an increasingly important method to explain the prevalence of cooperative behavior in human and other societies. In this paper, we have summarized results from two such experiments conducted in Beijing (Wu et al. 2009, 2012) and discussed these results in relation to theoretical predictions of their underlying games. For the two-player repeated PD game with costly punishment ("Costly Punishment and the PD Game" section), two unexpected experimental outcomes emerge; namely, that cooperation does not increase with the option of costly punishment and that a significant number of participants pre-emptively punish before observing whether their opponent is cooperating. These results are significantly different than those obtained in similar experiments conducted in Western societies. We conjecture that the differences may be attributed to cultural attitudes toward reputation. Specifically, building a good reputation in situations where the same two people interact repeatedly may be much less important in China than in the West and than in situations in China where several people interact repeatedly in a network.

For our PGG experiment with institutional reward and/or punishment, the incentive schemes are ordered as IRP, IP, IR, and then the Control in their effectiveness at promoting cooperation. This ordering is supported by the NE analysis of the underlying one-shot games in that Cooperate is the unique NE for IRP, IP has both Cooperate and Defect as its NE, and the only NE of IR is between these two strategies at a modest level of cooperation. The theoretical analysis can then be used to compare incentive schemes and their predicted effect on cooperative behavior.

Acknowledgments Support from the Natural Sciences and Engineering Research Council of Canada and from National Natural Science Foundation of China (Grant No. 31270439) is gratefully acknowledged.

References

Andreoni J, Harbaugh WT, Vesterlund L (2003) The carrot or the stick: rewards, punishments and cooperation. Am Econ Rev 93:893–902

Baldassarri D, Grossman G (2011) Centralized sanctioning and legitimate authority promote cooperation in humans. Proc Natl Acad Sci USA 108:11023–11027

Cressman R, Hofbauer J (2005) Measure dynamics on a one-dimensional continuous strategy space: theoretical foundations for adaptive dynamics. Theor Popul Biol 67:47–59

Cressman R, Song J, Zhang B, Tao Y (2012) Cooperation and evolutionary dynamics in the public goods game with institutional incentives. J Theor Biol 299:144–151

Dieckmann U, Law R (1996) The dynamical theory of coevolution: a derivation from stochastic ecological processes. J Math Biol 34:579–612

Dreber A, Rand DG, Fudenberg D, Nowak MA (2008) Winners don't punish. Nature 452:348–351

Falkinger J, Fehr E, Gächter S, Winter-Ebmer RA (2000) Simple mechanism for the efficient provision of public goods—experimental evidence. Am Econ Rev 90:247

Fehr E, Gächter S (2000) Cooperation and punishment in public goods experiments. Am Econ Rev 90:980–994

Grote D (2005) Forced ranking: making performance management work. Harvard Business School Press, Cambridge

Herrmann B, Thoni C, Gächter S (2008) Antisocial punishment across societies. Science 319:1362–1367

Isaac RM, Walker JM (1998) Communication and free-riding behavior: the voluntary contributions mechanism. Econ Inq 26:585–608

Ledyard J (1995) Public goods: a survey of experimental research. In: Kagel J, Roth A (eds) Handbook of experimental economics. Princeton University Press, Princeton

Mathew S, Boyd R (2011) Punishment sustains large-scale cooperation in prestate warfare. Proc Natl Acad Sci USA 108:11375–11380

Prendergast C, Topel R (1993) Discretion and bias in performance evaluation. Eur Econ Rev 37:355–365

Rand DG, Dreber A, Ellingsen T, Fudenberg D, Nowak MA (2009) Positive interactions promote public cooperation. Science 325:1272–1275

Rankin DJ, dos Santos M, Wedekind C (2009) The evolutionary significance of costly punishment is still to be demonstrated. Proc Natl Acad Sci USA 106:E135. doi:10.1073/pnas.0911990106

Schram A (2000) Sorting out the seeking: the economics of individual motivations. Public Choice 103:231–258

Sherratt TN, Wilkinson DM (2009) Big questions in ecology and evolution. Oxford University Press, Oxford

Sigmund K (2010) The calculus of selfishness. Princeton University Press, Princeton

Sigmund K, Hauert C, Nowak MA (2001) Reward and punishment. Proc Natl Acad Sci USA 98:10757–10762

Standifird S (2006) Using guanxi to establish reputation in China. Corp Reput Rev 9:171–178

Tao Y, Li C, Wu J, Cressman R (2009) The efficiency ratio of costly punishment. Proc Natl Acad Sci USA 106:E136. doi:10.1073/pnas.0912998107

Wu J, Zhang B, Zhou Z, He Q, Zheng X, Cressman R, Tao Y (2009) Costly punishment does not always increase cooperation. Proc Natl Acad Sci USA 106:17448–17451

Wu J, Li C, Zhang B, Xiao H, He Q, Cressman R, Tao Y (2012) Cooperation through institutional reward and punishment (mimeo)