Cooperation in Social Dilemmas: Free Riding May Be Thwarted by Second-Order Reward Rather Than by Punishment

Toko Kiyonari and Pat Barclay
McMaster University

Cooperation among nonrelatives can be puzzling because cooperation often involves incurring costs to confer benefits on unrelated others. Punishment of noncooperators can sustain otherwise fragile cooperation, but the provision of punishment suffers from a "second-order" free-riding problem because nonpunishers can free ride on the benefits from costly punishment provided by others. One suggested solution to this problem is second-order punishment of nonpunishers; more generally, the threat or promise of higher order sanctions might maintain the lower order sanctions that enforce cooperation in collective action problems. Here the authors report on 3 experiments testing people's willingness to provide second-order sanctions by having participants play a cooperative game with opportunities to punish and reward each other. The authors found that people supported those who rewarded cooperators either by rewarding them or by punishing nonrewarders, but people did not support those who punished noncooperators—they did not reward punishers or punish nonpunishers. Furthermore, people did not approve of punishers more than they did nonpunishers, even when nonpunishers were clearly unwilling to use sanctions to support cooperation. The results suggest that people will much more readily support positive sanctions than they will support negative sanctions.

Keywords: cooperation, punishment and reward, public goods game, social dilemma, altruism

In human societies, there are many examples of public goods problems, such as blood drives, payment for public radio programs, preservation of natural resources, and so on. The difficulty with providing such public goods is that by their very nature, no one can be denied access to them on the basis of whether he or she has contributed to their provision (Olson, 1965; Samuelson, 1954). As such, every individual has an incentive to undercontribute to the provision of public goods, but when everyone does this, everyone is worse off than they would have been had they all fully provided the public good. A classic example of a public good is the effort to protect the environment: Everyone is better off if everyone else cooperates by exercising restraint in resource use and investing in improving air and water quality, but each person would be individually best-off to personally use many resources and not invest in costly efforts in improving environmental quality. This can lead to possible tragedies of the commons (G. Hardin, 1968), where too little cooperation results in overuse or destruction of resources—global warming is one example of this process. Many academics will also recognize the public good of reviewing manuscripts for journals: Strong peer-reviewed science requires willing peers, but good reviews take personal time and effort. The problem of public-goods provision has been studied in many fields in both the natural and social sciences for decades (e.g., Bonacich, Shure, Kahan, & Meeker, 1976; Dawes, 1980; Fehr & Gächter, 2002; G. Hardin, 1968; R. Hardin, 1971; Olson, 1965; Ostrom, 1990), and the use of experimental games, which have long been popular in social psychology and other areas (e.g., Komorita & Parks, 1995; Messick & Brewer, 1983; Pruitt & Kimmel, 1977), is one methodology that bridges across these different disciplines. In

Toko Kiyonari and Pat Barclay, Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, Ontario, Canada.

Toko Kiyonari is now at the Department of Management, University of Antwerp, Antwerpen, Belgium; Pat Barclay is now at the Department of Neurobiology & Behavior, Cornell University.

We acknowledge support from a Social Sciences & Humanities Research Council of Canada grant to Margo Wilson and from a Natural Sciences & Engineering Research Council of Canada grant to Martin Daly. Toko Kiyonari was funded by a fellowship of the Japan Society for the Promotion of Science for conducting this research at McMaster University. We thank students at McMaster University for their participation. We gratefully thank Martin Daly and Margo Wilson for much advice, support, and comments on drafts of this article. We also thank Karthik Panchanathan, Toshio Yamagishi, and Matthijs van Veelen for comments on an earlier version of this article; Paul van Lange for comments on this version of the article; Shigeru Terai for assistance with programming; Trina Hancock, Abby Morrison, and Doug Vanderlaan for their assistance; and Andrew Clark, Greg Dingle, Danny Krupp, and Steve Stewart-Williams for their warm support, advice, and comments.

Correspondence concerning this article should be addressed to Toko Kiyonari, Department of Management, Stadscampus, Universiteit Antwerpen, S.Z. 501, Kipdorp 61, 2000 Antwerpen, Belgium. E-mail: Toko.Kiyonari@ua.ac.be or TokoKiyonari@gmail.com

Journal of Personality and Social Psychology, 2008, Vol. 95, No. 4, 826–842
 0022-3514/08/$12.00 DOI: 10.1037/a0011381

this article, we are focusing on the mechanisms that support human cooperative behavior through the lens of evolutionary psychology in order to see what factors support and maintain cooperation.

The Evolution of Human Cooperation and Punishment

Several theories have been proposed to explain the evolution of human cooperation, but they have difficulty explaining cooperation in large groups of nonrelatives. The mutual exchange of generous acts (i.e., reciprocal altruism; Trivers, 1971) explains how cooperation can evolve when individuals can target their beneficence preferentially toward cooperators and away from free riders who take the benefits of others' cooperation without paying the cost of reciprocating. This conditional cooperation (e.g., tit for tat; Axelrod, 1984) can thwart free riders in very small groups but not in larger groups of nonrelatives, such as in the provision of public goods, because one cannot selectively punish free riders by refusing further cooperation when such withholding hurts cooperators and free riders alike (Boyd & Richerson, 1988). Cooperation can also be stable when individuals carry a reputation for helping and if others are more likely to help those who have performed helping (indirect reciprocity; Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2003), but this also relies on people being able to help specific individuals and refuse help to noncooperators. When generosity can only be directed toward groups, such as in the provision of public goods, free riders cannot be excluded from benefiting, and such systems will not be stable.

However, cooperation can be maintained in groups when opportunities exist to selectively impose costs upon free riders (Fehr & Gächter, 2002; Ostrom, Walker, & Gardner, 1992; Yamagishi, 1986), and many recent studies have argued that such punishment has been crucial in the evolution of human cooperation (e.g., Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Gächter, 2002). Such punishment can include public criticism and disapproval, physical punishment, or monetary fines. People often impose costs upon those who violate norms of cooperation at a cost to themselves, and this costly punishment has been widely observed in both industrialized societies and small-scale societies (see Henrich et al., 2005). This fact, combined with neuropsychological evidence that people enjoy punishing noncooperators (de Quervain et al., 2004), suggests that a taste for punishment may be part of an evolved human psychology designed to maintain cooperation within large groups. People are certainly sensitive to social criticism and disapproval and are motivated to seek high self-esteem by maintaining positive evaluations of themselves (e.g., Allport, 1955; Baumeister, 1998; Jones, 1973; Sheldon, Elliot, Kim, & Kasser, 2001). However, if punishment is costly because of possible retaliation (e.g., Barclay, 2006; Cinyabuguma, Page, & Putterman, 2006), such as injury, energy expenditure, or reputational damage, then cooperators who do not punish will be better off than those who do. Thus, the provision of punishment (towards noncooperators) is itself a cooperative act that can be destabilized by others who free ride on these *moralistic* punishers (e.g., Boyd & Richerson, 1992; Henrich & Boyd, 2001; Oliver, 1980; Ostrom, 1990; Yamagishi, 1986), with some researchers even calling such punishment *altruistic* (from a functional perspective) due to the personal costs and the benefits to the group (Boyd et al., 2003; Fehr & Gächter, 2002; Fowler, 2005). Because of this personal cost, the existence of punishment itself requires explanation, such as a mechanism that would have selected for punitive sentiment in ancestral environments or social processes that cause punishing to be learned (or an interaction between the two).

Some theorists have suggested that punishment is maintained by the punishment of nonpunishers, which is sometimes called *second-order punishment* (e.g., Boyd & Richerson, 1992; Henrich, 2004; Henrich & Boyd, 2001) or *metanorms* (Axelrod, 1986), and other models on the evolution of punishment also assume the existence of such second-order punishment (e.g., Brandt, Hauert, & Sigmund, 2006; Fowler, 2005). Second-order punishment would be social disapproval or even physical punishment of those who fail to provide social disapproval and punishment of others who fail to cooperate—that is, if cooperation is a social norm, then people would not only punish those who break that norm by not cooperating (first-order punishment) but also provide second-order punishment toward those who break the metanorm of punishing norm-breakers (i.e., those who refuse to punish). As an example of second-order punishment, the United States has applied trade sanctions against nations and businesses that do not comply with sanctions against target countries that the United States has defined as uncooperative, using measures such as the Cuban Liberty and Democratic Solidarity Act of 1996 (a.k.a. the Helms-Burton Act), which imposes embargoes against foreign companies who trade with Cuba (U.S. Department of State International Information Programs, n.d.). In addition to demonstrating punishment of nonpunishers, this example demonstrates that the norms enforced by punishment and second-order punishment need not always be universally agreed upon as being good (Boyd & Richerson, 1992) and can be unrelated to cooperation, such as accusations of communism against those who did not denounce suspected communists in the McCarthy era in the United States, criminal charges against those who do not report serious crimes of which they are aware, or even punishment of those who did not support lynching in the American South (Axelrod, 1986). Second-order punishment may be more common when people have a vested interest due to alliances or competition, such as when people are wronged and get upset at friends or allies who befriend or otherwise do not punish the offenders (e.g., "you're either with us or against us"), because the second-order punishers then receive direct personal benefits for punishing. In cooperative situations like large-group public goods, these personal benefits are greatly diminished.

If second-order punishment is required to maintain first-order punishment of noncooperators, this argument would appear to be vulnerable to infinite regress (e.g., third-order punishment maintaining second-order punishment, etc.), but some theorists maintain that the demand for higher orders of punishment soon evaporates (see also Sober & Wilson, 1999). Once punishers and cooperators are both common, defection is not profitable, and although second-order free riders (who cooperate but do not punish) would have higher payoffs than would punishers because they avoid the costs of punishing, this advantage will be small if defection is a rare, costly mistake. As we ascend to higher orders of punishing, the demand for them becomes vanishingly small. Hence, it is argued, punishment is more likely to stabilize the punishment of nonpunishers than of noncooperators.

If this argument holds, then humans should possess a taste for second-order punishment, and it should be administered at least as readily, when the occasion calls for it, as first-order punishment.

But is this in fact how people behave? Kiyonari, van Veelen, and Yamagishi (2008) have shown that Japanese participants' willingness to administer second-order punishment was much lower than was their willingness to punish defectors, and they showed less approval for those who punished noncooperators than they did for those who abstained from punishing and instead free rode on the punishment provided by others. Seeing as second-order punishment of nonpunishers was not observed in that study, it seems that humans do not possess a readily elicited taste for this behavior. In contrast, it is fairly easy to elicit cooperation with strangers and (first-order) punishment of noncooperators, even in anonymous one-shot laboratory interactions (Fehr & Gächter, 2002), and tastes for such behaviors are said to be an evolved part of human psychology (Fehr & Fischbacher, 2003, 2004; Vogel, 2004).

Positive Sanctions and Negative Sanctions

Incentives can be either positive (rewarding cooperators) or negative (punishing defectors), and both can, in theory, maintain cooperation in collective actions (Panchanathan & Boyd, 2004). From a rational perspective, the distinction should not matter: Neither rewarding nor punishing is rational if providing these incentives is costly, and one can maximize one's gains by free riding on those provided by others (Oliver, 1980; Olson, 1965). However, we predicted that people would be more likely to administer second-order rewards (i.e., rewarding those who reward others) than they would second-order punishment for a combination of reasons.

Punishing nonpunishers (i.e., second-order punishment) is only evolutionarily stable when entire groups without punishers are more likely to go extinct or reproduce more slowly than groups with punishing and when these between-group selection pressures are sufficiently strong to overcome the individual disadvantages of punishing (Boyd et al., 2003; Henrich, 2004; Henrich & Boyd, 2001; Sober & Wilson, 1999). By contrast, selectively rewarding cooperators can be stable in collective action situations even without these assumptions because such selective rewards can be seen as part of a system of indirect reciprocity or generalized exchange when individuals help only those who have helped others (Panchanathan & Boyd, 2004). Under indirect reciprocity, people who help others gain a good reputation and are more likely to receive help from third parties, and such systems of indirect reciprocity are stable in at least some forms (Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2003). With public goods, people cannot selectively provide a public good toward cooperators and exclude free riders, but they may be able to selectively reward cooperators afterward. This first-order rewarding could start as a positive sanction, which could then become incorporated into a system of indirect reciprocity, where the second-order and higher order rewards are provided by others as part of that system: Cooperators receive more help than defectors do when in need (rewards), those who provide such help (rewarders) gain reputation and become more likely to receive help later (second-order rewards) than do those who refuse (nonrewarders), those who provide *that* help are more likely to receive help in turn, and so on. Panchanathan and Boyd (2004) have demonstrated that indirect reciprocity can support the provision of public goods, and Milinski and colleagues have shown that the presence of such reward systems can maintain contributions to public goods (e.g., Milinski, Semmann, & Krambeck, 2002; Rockenbach & Milinski, 2006), including in the fight against global warming (Milinski, Semmann, Krambeck, & Marotzke, 2006).

When punished, people respond with resentment or retaliation toward the punisher (e.g., Cinyabuguma et al., 2006), whereas they respond to helpful acts with increased empathy (e.g., Singer et al., 2006) and/or reciprocation. Thus, punishing incurs increased social costs, whereas rewarding brings increased benefits, so those who use rewards to motivate cooperation are likely to fare better than those who use punishment are. This is particularly true with second-order incentives because the distinction between moralistic second-order punishment and unprovoked aggression is unclear to witnesses unless they know the full sequence of prior behavior (see Barclay, 2006, for a discussion on responses to punishment that is or is not justified), whereas second-order rewards will simply seem like generosity to those who do not observe the prior sequence of behaviors. Furthermore, if the psychological cause of rewarding is a positive emotion, such as empathy (Batson et al., 1997) or concern for others, whereas the psychological cause of punishing is anger (e.g., Eisenberger, Lynch, Aselage, & Rohdieck, 2004; Fehr & Gächter, 2002), then people may be more willing to choose rewarders rather than punishers as social partners because they can benefit from associating with the former but be hurt by the latter. First- and second-order rewards could even constitute a form of competitive altruism (Barclay, 2004; Barclay & Willer, 2007; Hardy & Van Vugt, 2006; Roberts, 1998) if people are competing for access to social relationships.

For these reasons, we predicted that human psychology would be such that more people would be willing to provide second-order rewards than would provide second-order punishment and that people would be more supportive of the use of rewards than the use of punishment. We used a cooperative group game (public goods game [PGG]) to test these predictions. In the current article, Study 1 provided a first test of our ideas. In the second and third studies, we further explored the relation between punishment and reward by including different combinations of sanctions.

Testing the Product of Evolution, Not the Process of Evolution

Before we introduce our studies, let us justify an important feature of our research design, that is to say, the use of one-shot games instead of repeated games. We want to examine in these studies whether humans are readily elicited to punish nonpunishers (i.e., whether they have a taste or preference for such behavior), a hypothesis derived from evolutionary models (e.g., Boyd & Richerson, 1992) telling us that such a taste should have been selected through a long history of human evolution—definitely an iterated game environment. The fact that the particular psychology has evolved through repeated games does not mean that the selected psychology requires a repeated game environment to operate. Generally, selection takes place via the consequences of repeated applications of a particular strategy, but this does not necessarily imply that the selected strategy is conditional on the past sequences and/or future prospects. We have evolved social emotions, such as empathy and guilt, because they motivate social behavior that is probably adaptive on average in the "iterated game" of real life, but these emotions can be triggered even in one-shot anonymous laboratory settings where they may not be adaptive (Hagen

& Hammerstein, 2006). These preferences are informative, and because they are elicited so automatically, it suggests that they might be fast and frugal heuristics (Gigerenzer, 2001) that have evolved to solve problems in human social exchange (Cosmides & Tooby, 1992). What we are interested in here is whether humans are similarly predisposed to punish nonpunishers or to reward rewarders. One-shot games are more suitable for testing this than repeated games are because repeated games elicit strategic concerns as well as behavioral predispositions, and it would be impossible to tell whether any given behavior was produced by strategic thinking or by fast and frugal heuristics that were readily elicited. If second-order punishment or rewards are performed in one-shot games, it is less likely that these behaviors are caused by strategic concerns and more likely that people have a taste or preference for such behaviors. Repeated games would be suitable for testing the effects of strategic thinking on behavior (including the interaction between psychological inclinations and strategic thinking) or for testing the process through which the taste for punishment is selected (or acquired through learning), as is done in mathematical models and evolutionary simulations. In experiments, one tests whether people behave as predicted by the models, and this is our goal—to test whether people possess the predicted behavioral tendency to provide second-order punishment of nonpunishers.

Study 1

As a first test of our ideas, two conditions (punish vs. reward) were compared. First, our participants played a one-shot PGG without any information regarding subsequent opportunities for sanctions. Next, they were given opportunities to use either punishment or reward (i.e., first-order sanctions) without knowing about a subsequent second opportunity for sanctions (i.e., the second-order sanctions) in the following stage. This manipulation is appropriate because our question is whether or not sanctioning behavior can emerge at the first- and second-order levels without any shadow of the future. If an exogenous system of sanctions already exists, it makes cooperation a more explicit norm. People would possibly be more willing to follow the explicit rule and would want to avoid being punished (or want to receive more rewards), and also they may react more against noncooperators who ignore such an explicit social norm. However, our main focus here is on the psychological mechanisms underlying the emergence of sanctioning behavior endogenously or voluntarily, so we are not focusing on how people follow and enforce explicit social norms. Therefore, by using a one-shot game where participants did not expect either first- or second-order sanctioning opportunities, we can examine whether or not people take and respond to the cues of free riding voluntarily.

Method

Participants and anonymity. Ninety-seven 1st-year students at McMaster University played an anonymous one-shot PGG, described below, on computers in four-person groups. Confederates of the experimenter took the place of missing participants if one or more participants failed to arrive on time. Partitions prevented visual contact between participants, and communication was not permitted. All participants gave written, informed consent before participation and were debriefed upon completion of the experiment. Each participant was paid privately, and special care was taken to ensure the anonymity of participants' decisions and the privacy of the payment procedure. These methods were approved by the McMaster University Research Ethics Board.

PGG. In the PGG, each participant decided whether to contribute an endowment of C$5.00 (the conversion rate for Canadian dollars to U.S. dollars at the time was C$1 = US$0.85; all monetary values in this article are presented in Canadian dollars) to a public fund. Total contributions were to be doubled by the experimenter and divided equally among the four players, giving everyone a collective incentive to contribute but an individual incentive to keep their money. Before making any decisions, all participants had to successfully answer five questions that tested their understanding and ability to calculate payoffs. In fact, each participant unknowingly played the PGG against preprogrammed computer players and was told that one of the other three "players" had defected and two had cooperated.[1] All players were identified by code names only. The computer program was written in Visual Basic 6.0.

Punishment and reward. After the PGG contributions by each of the other code-named players were revealed, participants were given an unexpected opportunity to spend money to punish (punishment condition, 48 participants) or reward (reward condition, 49 participants) other players. They received three additional endowments of $1.00, any amount of which could be kept for themselves or spent to punish or reward each of the other three players (a total of $3.00) and were told that 3 times the amount they chose to spend would be removed from (punishment condition) or added to (reward condition) the target player's earnings. This multiplication meant that sanctions had a larger impact (positive or negative) on the recipient than they did on the participant administering them, which is a common assumption in studies of cooperation, punishment, and rewards. Punishment and reward were between-subjects conditions, and to reduce the effects of strategic planning (and therefore maximize measurement of psychological inclinations), participants were not forewarned about these opportunities to punish or reward. These surprise opportunities for sanctioning make the experiment an appropriate model of many real-life interactions, given that incomplete information is typical outside the laboratory, and people do not always know whether their actions are observed by people with whom they will later interact and who could sanction them.

After this first sanctioning stage, there was a second, again without prior warning (and participants were truthfully told that this second sanctioning stage was the final decision). Participants were told that one of the two cooperators in the PGG had spent either $0.80, $0.90, or $1.00 (randomly assigned with equal probability) to punish a defector (punishment condition) or to reward the other cooperator (reward condition) and that the other two

[1] We believe that deception was necessary in our studies to control the situation such that participants would face a sanctioner and a nonsanctioner while controlling for group levels of cooperation and individual differences in rewarding and punishing. Without this manipulation, many participants (~70%) would not simultaneously encounter a cooperative sanctioner and nonsanctioner (for example, due to too many/few other defectors/cooperators in a group or too many/few other sanctioners).

players (one cooperator and one defector) had spent nothing on punishment or rewards, but they were not yet informed whether they themselves had received sanctions. After receiving this information about others' sanctioning decisions, participants received another endowment of \$1.00 to punish/reward each of the three players. The other three players in the punishment condition were (a) the punisher, who had cooperated in the PGG and punished the defector in the first stage of sanctions; (b) the nonpunisher, who had also cooperated in the PGG but had not punished the defector; and (c) the defector, who had not contributed in the PGG and did not have a target of punishment in the first stage of sanctions because he/she was the only PGG defector. The focal comparison was whether more punishment would be administered in the final round to the nonpunisher or to the punisher. Similarly, the other three players in the reward condition were (a) the rewarder, who cooperated in the PGG and rewarded the other cooperator in the first sanctioning state; (b) the nonrewarder, who cooperated but did not reward player (a); and (c) the defector, who neither cooperated in the PGG nor rewarded anyone. The focal comparison was whether more reward would be given to the rewarder or to the nonrewarder. Figure 1 shows the flow of Study 1 from each participant's perspective.

Postexperimental questionnaire. Following the second incentive stage, participants completed a postexperimental questionnaire in which they used a 9-point scale (anchored at 1 = *strongly disagree* and 9 = *strongly agree*) to rate the other players (by code name) on trustworthiness, cooperativeness, generosity, likability, goodness, and dependability. These six items were highly correlated (Cronbach's alphas $> .82$), so we combined them into a single measure of favorable evaluation.

Statistical analyses. Because each participant encountered a sanctioner and a nonsanctioner, within-subjects analyses were generally called for. We used a repeated measures analysis with weighted least squares to analyze the categorical data on the number of participants who sanctioned each type and a Wilcoxon signed-ranks test to analyze the ratio data on the amount of money spent to sanction each type. For between-subjects comparisons across experimental conditions or for different types of participants, we used Wilcoxon rank-sum tests. To simplify the presentation of the results, we will simply present the statistics (z values for Wilcoxon rank-sum tests and S values for Wilcoxon signed-ranks tests) rather than naming the tests each time.

Results

Cooperation and first-order sanctions. The punishment and reward conditions did not differ in cooperation levels in the contributions stage, 65% (31/48) versus 63% (31/49), respectively. In the punishment condition, more participants administered first-order punishment to the defector than to one or both of the cooperators (23/48 vs. 9/48, respectively), $\chi^2(1, N = 48) = 9.84$, $p = .002$, and more was spent on first-order punishment of defectors than was spent on cooperators ($S = 155$, $p < .001$; see Figure 2A). In the reward condition, more participants provided first-order rewards to the cooperators than to the defector (40/49 vs. 10/49, respectively), $\chi^2(1, N = 49) = 51.04$, $p < .001$, and

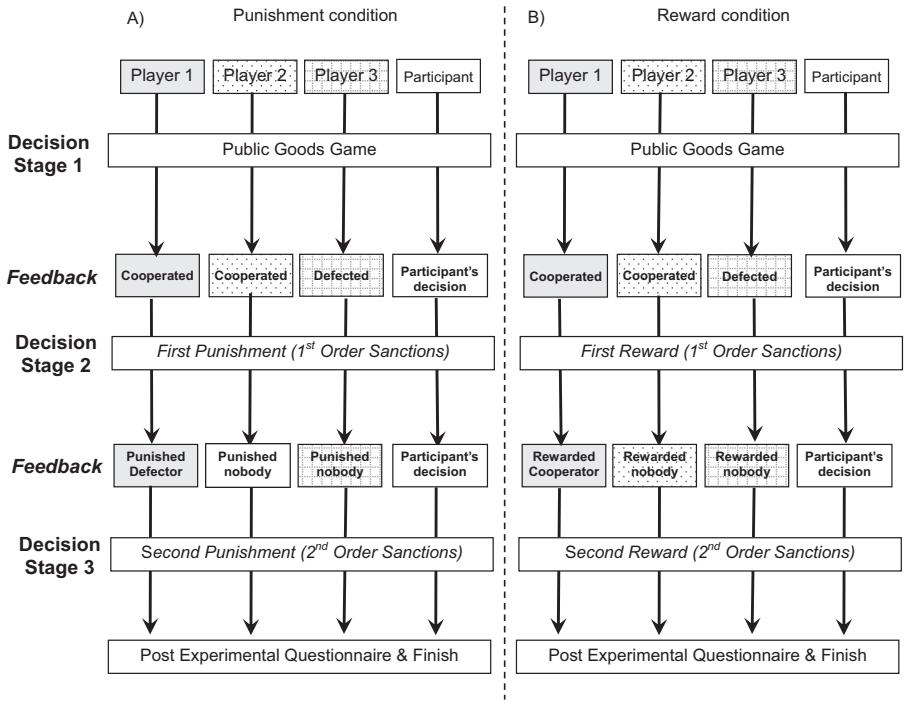

Figure 1. Flow of Study 1 from each participant's perspective in the punishment (A) and reward (B) conditions. For the sake of clarity in this figure, Player 1 is always a cooperator/sanctioner, Player 2 is always a cooperator/nonsanctioner, and Player 3 is always a defector/nonsanctioner, but in the real experiment, the player numbers associated with each other "person" could vary.

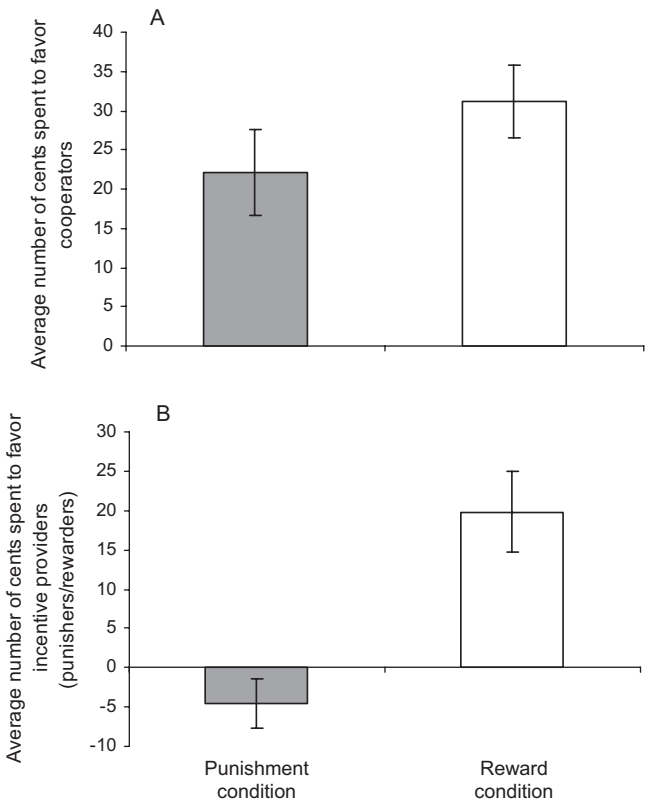

Figure 2. Differential sanctioning favoring cooperators (A) and incentive providers (B) in Study 1. Positive amounts indicate favorable treatment (more money spent on reward or less spent on punishment) of cooperators over defectors and incentive providers over nonproviders. Error bars represent standard errors of the differences in amounts spent on incentives. A: At the first opportunity to provide incentives, defectors received more punishment and fewer rewards than did cooperators in both the reward and punishment conditions (Wilcoxon signed-ranks test: $S = 345$ and $S = 155$, respectively, both ps $< .001$). B: Participants did not spend more to punish nonpunishers than they spent on punishers at the second opportunity for incentives (Wilcoxon signed-ranks test: $S = 4$, $p = .250$) but did spend more to reward rewarders than they spent on nonrewarders (Wilcoxon signed-ranks test: $S = 76.5$, $p < .001$).

more was spent on first-order rewards to each cooperator than was spent on the defectors ($S = 345$, $p < .001$; see Figure 2A). More money was spent to reward each cooperator than was spent to punish each defector (39.5¢ vs. 26.9¢, respectively; $z = 2.56$, $p = .010$). Cooperators fared slightly better relative to defectors in the reward condition than they did relative to defectors in the punishment condition, but this difference was nonsignificant ($z = 1.78$, $p = .075$; see Figure 2A).

We classified participants according to whether they had themselves cooperated in the PGG. In the punishment condition, 23.5% (4/17) of noncooperative participants spent some money to provide first-order punishment to the defector, whereas 61.3% (19/31) of cooperative subjects did so ($z = 2.62$, $p = .009$), and the two types of participants differed in the degree to which they punished defectors more than they punished cooperators (average differences: 1.5¢ vs. 33.5¢, respectively; $z = 2.95$, $p = .003$). In the reward condition, 88.9% (16/18) of noncooperative participants and 83.9% (26/31) of the cooperative participants spent some money to provide first-order rewards to the cooperators ($z = 0.69$, $p = .49$), and the two types of participants did not differ in the degree to which they rewarded cooperators more than they rewarded defectors (average differences: 28.6¢ vs. 32.6¢, respectively; $z = 0.37$, $p = .71$). Thus, only cooperative participants were likely to punish defection, but cooperators and noncooperators both rewarded cooperation.

Second-order sanctions. At the second sanctioning stage, more was spent on rewards than was spent on punishment (47.6¢ vs. 19.8¢, respectively; $z = 2.99$, $p = .003$). Only 1 of 48 participants in the punishment condition punished the cooperator/nonpunisher more than he/she did the cooperator/punisher, whereas 3 of 48 punished the cooperator/punisher more than they did the cooperator/nonpunisher, and 2 punished them equally. As a result, slightly (but not significantly) more money was spent to punish the punisher than was spent to punish the nonpunisher ($S = 4$, $p = .25$; see Figure 2B), a difference directionally opposite to what should have been observed were people inclined to administer the sort of second-order punishment that some theories demand. In the reward condition, by contrast, second-order sanctions of the sort required to support cooperation were provided: More participants rewarded the rewarder than the nonrewarder (30/49 vs. 16/49, respectively), $\chi^2(1, N = 49) = 16.33$, $p < .001$, and more was spent on rewarders ($S = 76.5$, $p < .001$; see Figure 2B). The advantage of rewarders relative to nonrewarders was greater than was the advantage of punishers relative to nonpunishers (which was actually a disadvantage)—that is, there was more second-order rewarding than there was second-order punishment ($z = 4.06$, $p < .001$).

Cooperative participants punished the nonpunishers slightly but not significantly more than they did the punishers (1.3¢ vs. 0.3¢, respectively; $S = 0.5$, $p = 1.00$), whereas noncooperative participants spent slightly more to punish the punishers than they did to punish the nonpunishers (15.9¢ vs. 1.2¢, respectively; $S = 3$, $p = .25$), and these two patterns of punishment differed significantly ($z = 2.38$, $p = .017$). In contrast, the rewarders were rewarded more than the nonrewarders were by both cooperative participants (42.6¢ vs. 17.4¢, respectively; $S = 46$, $p = .002$) and noncooperative participants (13.3¢ vs. 2.8¢, respectively; $S = 5$, $p = .13$), and the two participant types did not differ in their degree of favoritism toward rewarders (25.2¢ vs. 10.6¢, respectively; $z = 1.30$, $p = .20$).

Postexperimental questionnaires. Cooperators were evaluated more favorably than defectors were in both the punishment and reward conditions (average differences: 2.97 and 3.12, respectively; $S = 528$ and 571, respectively, ps $< .001$; see Figure 3A). In the reward condition, cooperative participants did not differ from noncooperative participants in their ratings of the cooperators relative to defectors (average differences were 3.47 vs. 2.53, respectively; $z = 1.48$, $p = .14$), so we are justified in combining all participants in the reward condition into one analysis. In the punishment condition, cooperative participants showed a greater preference for the cooperators than did noncooperative participants (average differences in ratings of cooperators and defectors: 3.78 and 1.48, respectively; $z = 3.13$, $p = .002$), but both types of participants evaluated cooperators more positively than they did defectors ($S = 229$, $p < .0001$, and $S = 57.5$, $p < .005$, respectively).

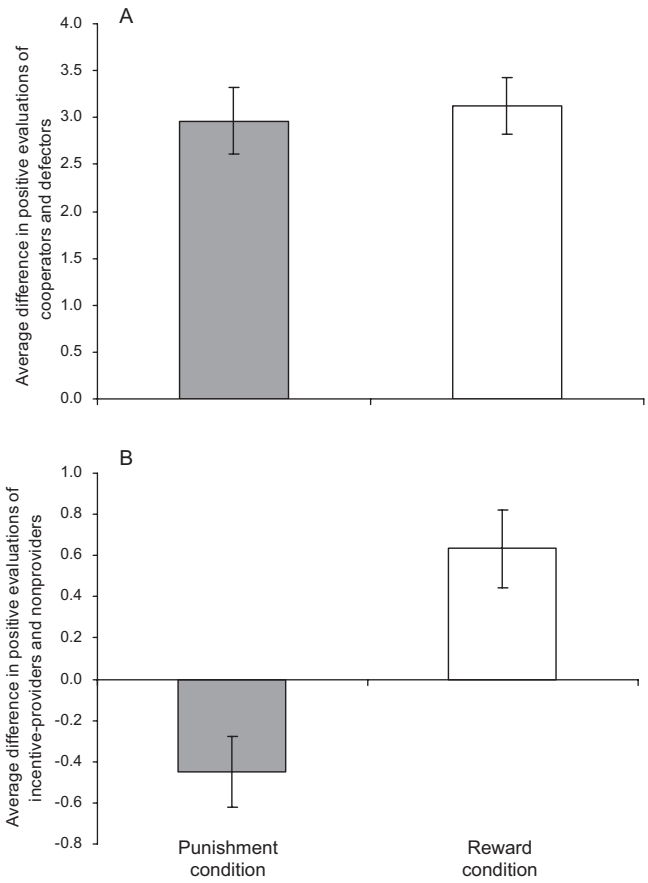

Figure 3. Positive evaluations in Study 1. The graph shows the average difference in positive evaluations between cooperators and defectors (A) and between incentive providers and nonproviders (B). Positive ratings were a composite of ratings of trustworthiness, cooperativeness, generosity, likeability, goodness, and dependability. Error bars represent the standard error of the difference in evaluations. A: Cooperators were rated more positively than were defectors in both the punishment and reward conditions (Wilcoxon signed-ranks test: $S = 528$ and $S = 571$, respectively, both $ps < .001$). B: Punishers were rated less positively than were nonpunishers (Wilcoxon signed-ranks test: $S = 148.5$, $p = .023$), but rewarders were rated more positively than were nonrewarders (Wilcoxon signed-ranks test: $S = 203.5$, $p = .001$).

Punishers were evaluated less favorably than were nonpunishers (average difference in ratings: −0.45; see Figure 3B) in this one-round game ($S = 148.5$, $p = .023$), and cooperative and uncooperative participants did not differ in their ratings of punishers relative to their ratings of nonpunishers (mean differences in ratings between punishers and nonpunishers: −0.35 and −0.63, respectively; $z = 0.86$, $p = .39$). This could explain why participants did not practice second-order punishment: They did not regard punishment—even punishment for failing to provide a public good—as a socially desirable act. In sharp contrast, participants rated rewarders more favorably than they did nonrewarders (mean difference in ratings between rewarders and nonrewarders: 0.63; $S = 203.5$, $p = .001$; see Figure 3B), and cooperative participants did not differ from noncooperative participants in their ratings of rewarders relative to their ratings of nonrewarders (mean differences in ratings: 0.79 and 0.36, respectively; $z = 0.86$, $p = .39$). Thus, it appears that cooperative and uncooperative participants show similar patterns of ratings toward cooperators versus defectors, punishers versus nonpunishers, and rewarders versus nonrewarders. Rewarders were rated more positively relative to nonrewarders than punishers were relative to nonpunishers ($z = 3.51$, $p < .001$; see Figure 3B), even though rewarders and punishers both provided costly incentives for cooperation.

Discussion

Our findings replicated previous research that has shown that people will administer costly first-order punishment (Fehr & Gächter, 2002; Kiyonari et al., 2008; Ostrom et al. 1992; Yamagishi, 1986) and first-order rewards (Milinski et al., 2002). However, our participants did not provide significant second-order punishment of nonpunishers nor did they perceive first-order punishment positively, which speaks against models that rely on the existence of second-order punishment (Fowler, 2005; Henrich, 2004; Henrich & Boyd, 2001). This interpretation is supported by the results of Kiyonari et al. (2008), whose Japanese participants also failed to provide second-order punishment. Moreover, Barclay (2006) found no second-order punishment in PGGs with multiple rounds.

Unlike second-order punishment, rewarding cooperators was viewed as more socially desirable than not rewarding was by cooperative and uncooperative participants, and both types of participants were willing to perform selective second-order rewarding such that the rewarders actually received more money from group members than the nonrewarders did. This is consistent with the idea that rewarding cooperation is part of a self-sustaining system of indirect reciprocity (Panchanathan & Boyd, 2004). These results are strengthened by the fact that cooperative and noncooperative participants did not differ in their selective rewarding of cooperators and rewarders, and even noncooperative participants (who should have had no expectation of receiving rewards) rated rewarders favorably relative to nonrewarders. Thus, theories that consider the administration of punishment and of reward equivalent in their impacts on the stability of cooperation (e.g., see review in Oliver, 1980) do not correspond to how people perceive and react to these incentives.

To conclude Study 1, second-order punishment is not a likely candidate for the solution of the second-order free-riding problem because (a) punishing defectors was not regarded as socially desirable in this one-shot PGG, and thus (b) people did not administer second-order punishment. However, rewarding rewarders was considered socially desirable, and people did provide these second-order rewards. The latter conclusion is consistent with the recent model by Panchanathan and Boyd (2004) on indirect reciprocity that shows that it is evolutionarily stable to discriminate against those who do not reward cooperators in collective action problems. In this model, rewarding is viewed as cooperation at any order—that is, first-order rewarding is as desirable as cooperation is in the original public goods provision, second-order rewarding is as desirable as first-order rewarding, ad infinitum. Cooperation in collective actions may be treated the same as any other cooperative act in ongoing systems of indirect reciprocity in humans. Thus, rewarding is not vulnerable to the problem of second-order free riding if rewards are part of such a system of generalized exchange

because people must reward in order to remain in good standing in their group. The results of our postexperimental questionnaire suggest that this is a good candidate for solving the puzzle of how cooperation among nonrelatives can be maintained; the desirability of cooperation does not drop with the provision of rewards as it does with the provision of punishment. This makes rewards a more likely candidate than punishment is for the maintenance of large-scale human cooperation.

However, the results of Study 1 allow for a possibility that punishment may be supported by second-order rewarding (rather than second-order punishment). Further, they are silent about the question of whether rewarding is or is not supported by second-order punishment (rather than second-order rewarding). People might praise punishers by rewarding them. They might also disapprove of nonrewarders by punishing them at the second-order sanction level. Until we test these possibilities, we cannot firmly conclude that punishment is less likely than rewards are to have been a factor in the emergence of large-scale human cooperation. We thus conducted a second study to explore these possibilities. Specifically, we address the following two questions in the second study: whether people reward punishers and whether they punish nonrewarders. Answering these questions will provide more definite answers as to whether people endorse punishment and/or rewards as a means for promoting cooperation.

Study 2

The design of the second study was almost identical to that of the first study, except for the types of sanctions participants were provided with at the two levels. In Study 1, participants could provide the same type of sanctions at the first and second levels. In Study 2, the two types of sanctions were crossed. In the punish–reward condition, participants were given an opportunity to punish cooperators and/or noncooperators at the first level and then to reward punishers and/or nonpunishers at the second level. In the reward–punishment condition, the orders of the two types of sanctions were reversed. The rest of the procedures were exactly the same as in Study 1. To avoid repetition, we describe the procedure only briefly.

Method

Participants were 80 1st-year students at McMaster University. Participants played a one-shot anonymous PGG in four-person groups and received an unexpected opportunity to spend money to punish (punish–reward condition, 40 participants) or reward (reward–punish condition, 40 participants) other players. After this first sanctioning stage, they were told that one of the two cooperators in the PGG had spent money to punish a defector (punish–reward condition) or to reward the other cooperator (reward–punish condition) and that the other two players (one cooperator and one defector) had spent nothing on punishment or rewards. As in Study 1, the participants were not informed whether they had received rewards or punishment to eliminate the possibility of directly reciprocal behavior toward other players. After receiving this information about others' decisions, participants received another endowment of money to reward (punish–reward condition) or punish (reward–punish condition) each of the three players. The endowments and the structure for the PGG and the sanctioning were the same as in Study 1. The punish–reward and reward–punish conditions were a between-subjects factor, and participants were not forewarned about these opportunities to punish or reward.

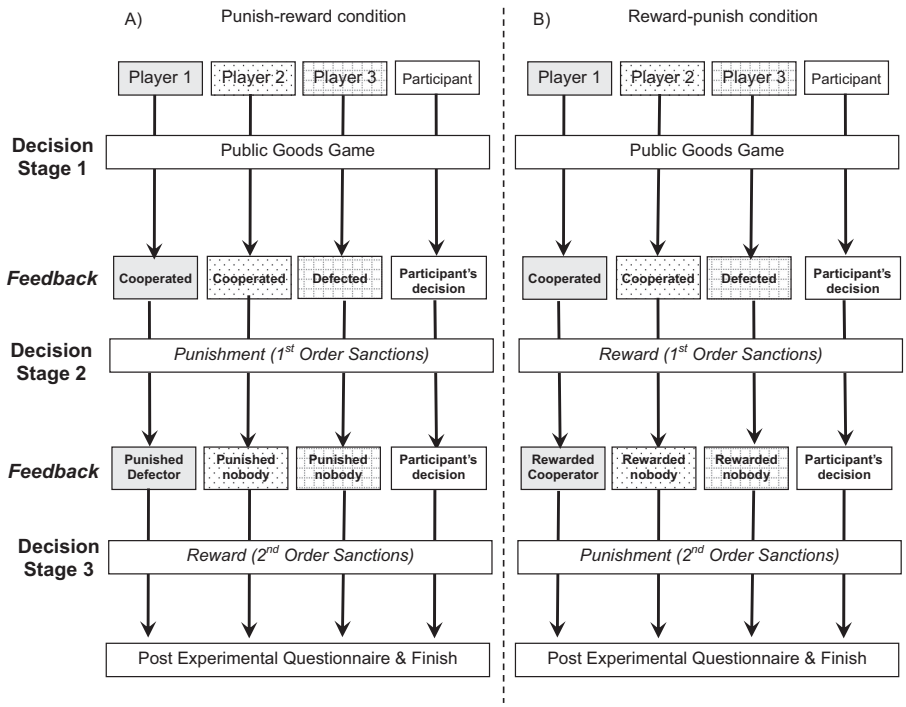

Figure 4. Flow of Study 2 from each participant's perspective in the punish–reward (A) and reward–punish (B) conditions. As in Study 1, the player numbers associated with each other "person" could vary.

Following the second sanction stage, participants completed a postexperimental questionnaire that included the six evaluation items used in Study 1, and we combined them into a single measure of favorable evaluation, as in Study 1. Figure 4 shows the flow of Study 2.

Results

Cooperation and first-order sanctions. The punish–reward and reward–punish conditions did not differ in cooperation levels in the contribution stage, 65% (26/40) versus 67.5% (27/40), respectively. In the punish–reward condition, more participants administered first-order punishment to the defector than to one or both of the cooperators (18/40 vs. 4/40, respectively), $\chi^2(1, N = 40) = 17.66$, $p < .0001$, and more was spent on first-order punishment of defectors (difference was 22.5¢; $S = 94$, $p < .0001$; see Figure 5A). In the reward–punish condition, more participants provided first-order rewards to the cooperators than to the defector (35/40 vs. 8/40, respectively), $\chi^2(1, N = 40) = 83.08$, $p < .0001$, and more was spent on first-order rewards to cooperators (difference was 28.3¢; $S = 248$, $p < .0001$; see Figure 5A). More money was spent to reward each cooperator than was spent to punish each defector (32.5¢ vs. 23.8¢, respectively; $z = 2.18$, $p = .029$). Cooperators fared slightly better (i.e., received more reward or less punishment) relative to defectors in the reward–punish condition (28.3¢) than they did relative to defectors in the punish–reward condition (22.5¢), but this difference was nonsignificant ($z = 1.70$, $p = .089$).

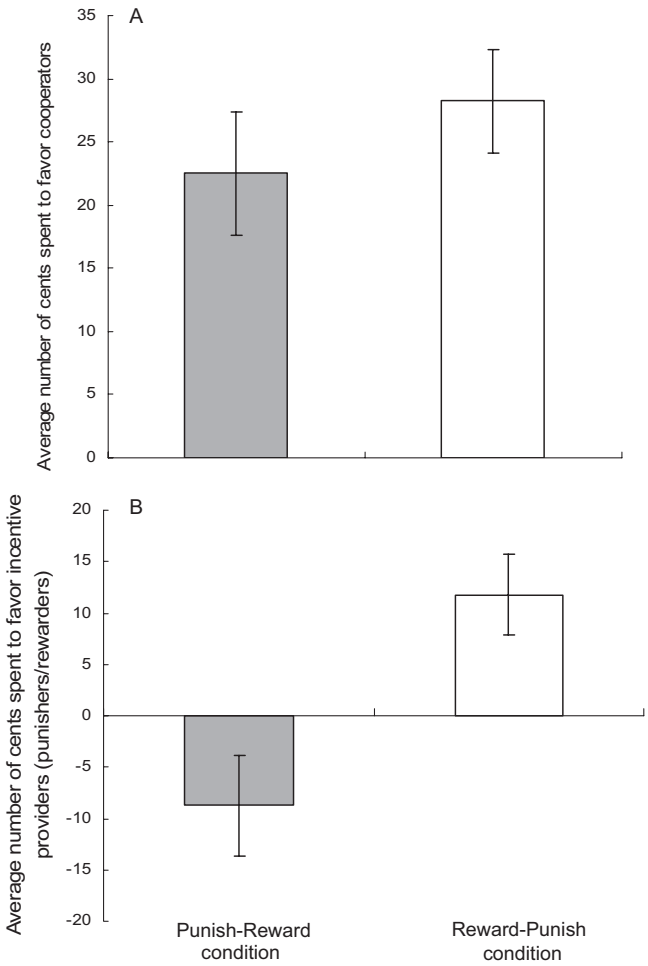

Figure 5. Differential sanctioning favoring cooperators (A) and incentive-providers (B) in Study 2. Error bars represent the standard error of the difference in amounts spent on incentives. A: At the first opportunity to provide incentives, defectors received more punishment and fewer rewards than did cooperators in both the punish–reward and reward–punish conditions. B: Participants did not spend more to reward punishers than they spent on nonpunishers at the second opportunity for incentives in the punish–reward condition but did spend more money to punish nonrewarders than they spent on rewarders in the reward–punish condition.

We classified participants according to whether or not they had themselves cooperated in the PGG. In the punish–reward condition, 21.4% (3/14) of noncooperative participants spent some money to provide first-order punishment to the defector, whereas 57.7% (15/26) of cooperative subjects did ($z = 1.91$, $p = .056$), and the two types of participants differed in the degree to which they punished defectors more than they punished cooperators (average differences: 8.93¢ vs. 29.8¢, respectively; $z = 2.32$, $p = .02$). In the reward–punish condition, 84.6% (11/13) of noncooperative participants and 88.9% (24/27) of the cooperative participants spent some money to reward the cooperators ($z = 0.13$, $p = .90$), and statistically the two types did not differ in the degree to which they rewarded cooperators more than they rewarded defectors (average differences: 21.2¢ vs. 31.7¢, respectively; $z = 1.14$, $p = .26$). Thus, only cooperative participants were likely to punish defection, but cooperators and noncooperators both rewarded cooperation. These results replicate the results of the first-order sanctions in Study 1.

Second-order sanctions. At the second sanctioning stage in the punish–reward condition, 21 of 40 participants (52.5%) rewarded the cooperator/punisher and 26 of 40 participants (65.0%) rewarded the cooperator/nonpunisher, though this was not statistically significant, $\chi^2(1, N = 40) = 2.41$, $p = .12$. More money for second-order rewards was spent on nonpunishers than was spent on punishers, though this was nonsignificant (28.3¢ vs. 19.5¢, respectively; $S = 21$, $p = .099$; Figure 5B). These results show that subjects did not provide second-order rewarding that could support moralistic punishers because punishers did not receive more rewards (and in fact, received fewer) than nonpunishers did.

However, in the reward–punishment condition, 3 of 40 participants (7.5%) punished the cooperator/rewarder, and 10 of 40 participants (25%) punished the cooperator/nonrewarder, $\chi^2(1, N = 40) = 6.30$, $p = .012$. More money for punishment was spent on the nonrewarder than was spent on the rewarder (12.8¢ vs. 1.0¢, respectively; $S = 31.5$, $p = .003$; see Figure 5B). Therefore, subjects provided second-order punishment that could support moralistic rewarders because people punished those who failed to reward cooperators more than they did those who did reward. Then the advantage of rewarders relative to nonrewarders (11.8¢) was greater than was the advantage of punishers relative to nonpunishers (which was actually a disadvantage, −8.8¢; $z = 2.80$, $p = .005$). This result shows that moralistic rewarding behavior toward cooperators could be supported by either subsequent punishment (Study 2) or reward (Study 1), but moralistic punishment behavior toward noncooperators was supported by neither reward (Study 2) nor punishment (Study 1).

Both cooperative participants and noncooperative participants showed the same direction in behavioral patterns of second-order sanctions in both conditions, although the degree of strength differed. In the punish–reward condition, cooperative participants rewarded the nonpunisher slightly but not significantly more than they did the punisher (31.2¢ vs. 28.1¢, respectively; $S = 2.5$, $p = .688$), whereas noncooperative participants spent more (but not quite significantly more) to reward the nonpunishers than they did the punishers (22.9¢ vs. 3.6¢, respectively; $S = 8.5$, $p = .094$). These two patterns of rewarding did not differ significantly ($z = 1.61$, $p = .11$), and neither type of participant provided second-order rewarding of punishers. In the reward–punish condition, cooperative participants punished nonrewarders more strongly than they did rewarders (16.7¢ vs. 1.1¢, respectively; $S = 22.5$, $p = .0039$), and noncooperative participants punished nonrewarders slightly but not significantly more than they did rewarders (4.6¢ vs. 0.8¢, respectively; $S = 0.5$, $p = 1.00$). Although the difference between these two participant types approached significance in their tendency to punish nonrewarders ($z = 1.91$, $p = .056$), both types of participants did provide some second-stage selective punishment of nonrewarders. To summarize, both cooperative and noncooperative participants tended to respond similarly in both conditions; however, cooperative participants showed stronger support of rewarders than did noncooperative participants in the reward–punishment condition (i.e., second-order punishment), whereas noncooperative participants showed stronger support of nonpunishers than did cooperative participants in the punish–reward condition (i.e., "reversed" second-order rewarding).

Postexperimental questionnaires. Cooperators were evaluated more favorably than defectors were in both the punish–reward condition and the reward–punish condition (average differences: 2.36 vs. 3.46, respectively; $S = 369$ and 390, respectively, ps $< .0001$; see Figure 6A). In the punish–reward condition, cooperative participants did not differ from noncooperative participants in their ratings of the cooperators relative to defectors (average differences are 2.65 vs. 1.82, respectively; $z = 0.74$, $p = .461$), so we are justified in combining all participants in the punish–reward condition into one analysis. In the reward–punish condition, cooperative participants showed more of a preference for the cooperators than did noncooperative participants, but the degree of difference was nonsignificant (average differences in ratings of cooperators and defectors: 3.86 vs. 2.63, respectively; $z = 1.82$, $p = .068$), and both types of participants evaluated cooperators more positively than they did defectors ($S = 189$ and 39, respectively, ps $< .001$). It seems that people definitely like cooperators in one-shot PGGs more than they do noncooperators, regardless of what type of sanctions they experienced.

In the punish–reward condition, punishers were evaluated slightly less favorably than nonpunishers were (average difference in ratings: −0.31; see Figure 6B), but this difference was not significant ($S = 62$, $p = .32$). Cooperative participants did not discriminate between punishers and nonpunishers, but noncooperative participants evaluated punishers more negatively than they did nonpunishers; this difference in ratings between cooperative and noncooperative participants was nonsignificant (mean differences in ratings between punishers and nonpunishers: 0.04 and −0.98, respectively; $z = 1.66$, $p = .096$). This shows that participants did not evaluate punishers positively relative to nonpunishers regardless of the type of sanctions they faced at the second stage (punishment in punishment condition in Study 1, reward in punish–reward condition in Study 2).

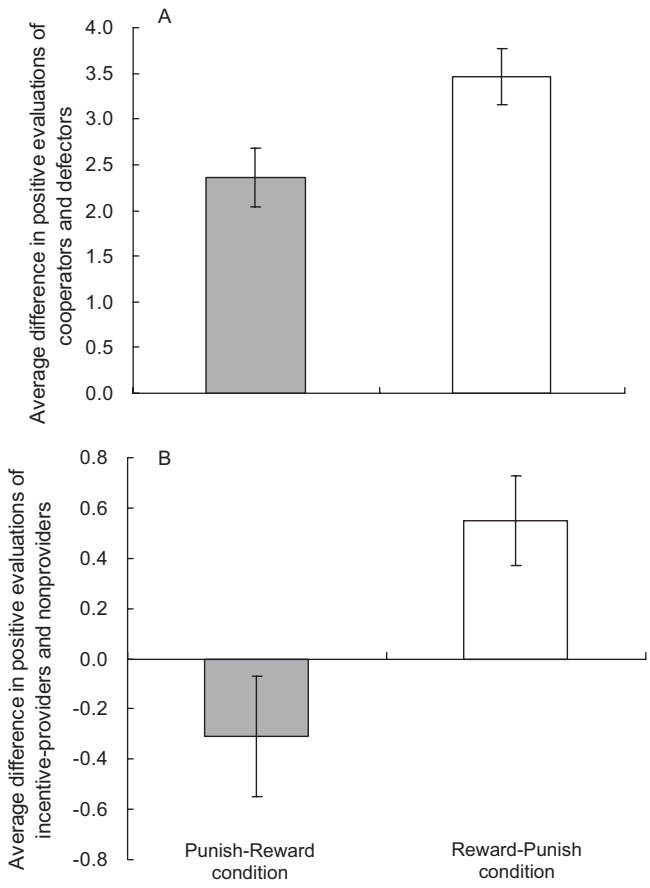

Figure 6. Positive evaluations in Study 2. The graph shows the average difference in positive evaluations between cooperators and defectors (A) and between incentive providers and nonproviders (B). Error bars represent the standard error of the difference in evaluations. A: Cooperators were rated more positively than defectors were in both the punish–reward and reward–punish conditions. B: Punishers were rated less positively than were nonpunishers, but rewarders were rated more positively than were nonrewarders.

In sharp contrast, in the reward–punish condition, participants rated rewarders more favorably than they did nonrewarders (mean difference in rating between rewarders and nonrewarders: 0.55; $S = 122$, $p = .0095$; see Figure 6B), and cooperative participants did not differ from noncooperative participants in their ratings of rewarders relative to nonrewarders (mean differences in ratings: 0.62 and 0.41, respectively; $z = 0.34$, $p = .74$). This pattern was the same as that in Study 1.

Study 2's results indicate that cooperative and noncooperative participants show similar patterns of ratings toward cooperators versus defectors and rewarders versus nonrewarders. They differ slightly in their ratings toward punishers and nonpunishers, although neither type of participant had a significant preference for punishers over nonpunishers. Rewarders were rated more positively (relative to nonrewarders) than punishers were (relative to

nonpunishers; $z = 2.27$, $p = .023$), even though rewarders and punishers both provided costly incentives for cooperation.

Summary of Study 2

Study 2 replicated the pattern of the first-order sanctions that we observed in Study 1. As in Study 1, more money was spent to reward cooperators than was spent to punish defectors. Furthermore, Study 2 shows that rewarding cooperators is supported by second-order punishment, but punishing defectors is not supported by second-order reward. This result is consistent with the result of Study 1 in the sense that participants respond to the target person's rewarding behavior (or lack of it) by either rewarding rewarders (Study 1) or punishing nonrewarders (Study 2), whereas they are not responsive to the target person's punishment behavior (or lack of it) either in the form of punishing nonpunishers (Study 1) or rewarding punishers (Study 2). Our participants did not seem to care whether another participant provided punishment or not, whereas they did care whether another participant provided rewards or not. Both rewarding cooperators and punishing noncooperators are costly sanctions for the maintenance of a public good. Despite this fact, there appears to be an asymmetry in participants' sensitivity toward the first-order sanctions.

There is still the possibility that this asymmetry is caused by ambiguity about the intentions of the punishers—"justified" punishment of free riders can be hard to distinguish from "unjustified" spiteful or aggressive actions, and nonpunishment due to mercy is hard to discriminate from deliberate second-order free riding (Barclay, 2006). This ambiguity can be reduced if participants have opportunities to reward cooperators and punish free riders at the same time: If someone punishes free riders *and* rewards cooperators, then the punishment is more likely to be norm-enforcing punishment rather than spiteful punishment. Similarly, if someone does neither, then that lack of sanctions is more likely to be second-order free riding than it is to be mercy toward noncooperators, and we might expect disapproval of that type of nonpunishment. To explore these possible effects of ambiguity of punishers' and nonpunishers' intention, we conducted a third experiment, in which participants can choose either punishment or reward at the same time.

Study 3

Introduction and Method

The design of this experiment was almost identical to that of the previous studies (Study 1 and Study 2) except in the types of sanctions. In Studies 1 and 2, participants were given only one type of sanction at each sanction stage; in Study 3, both types of sanctions (punishment and reward) were available at each sanctioning stage. Participants always encountered one defector and two cooperators, one of whom never rewarded or punished anyone (nonsanctioner) and the other of whom was one of three types of sanctioner: a punisher, one who only punished the defector (punisher condition, 41 participants); a rewarder who only rewarded the other cooperator (rewarder condition, 37 participants); or a fair sanctioner who punished the defector and rewarded the other cooperator (fair-sanctioner condition, 38 participants). The type of sanctioner experienced was a between-subjects factor (total $N = 116$). This design allowed us to test whether fair sanctioners (who perform both types of sanctions) receive more approval than other types of sanctioners do. Figure 7 shows the flow of Study 3.

Results

Cooperation and first-order sanctions. The three experimental conditions did not differ until after the first sanctioning stage, so they should have similar levels of cooperation in the PGG and of sanctioning at the first sanction stage. Indeed, they did not differ in cooperation levels in the contribution stage, 58.5% (24/41) in the punisher condition versus 67.8% (25/37) in the rewarder condition versus 65.8% (25/38) in the fair-sanctioner condition, $\chi^2(1, N = 116) = 0.78$, $p = .68$.

In all three conditions, more participants administered first-order punishment to the defector than to one or both of the cooperators (17/41 vs. 3/41 in the punisher condition; 22/37 vs. 2/37 in the rewarder condition; 18/38 vs. 5/38 in the fair-sanctioner condition, respectively), and this was highly significant using a repeated measures analysis with weighted least squares, with the presence of punishment toward the defector and the cooperators as dependent variables, $\chi^2(1, N = 116) = 60.12$, $p < .0001$. This analysis revealed no differences between the three experimental conditions, $\chi^2(2, N = 116) = 1.69$, $p = .43$, and no significant interaction between experimental condition and target of punishment, $\chi^2(2, N = 116) = 3.24$, $p = .20$. More money was spent on providing first-order punishment toward defectors than was spent on cooperators in all three conditions (for the punisher, rewarder, and fair-sanctioner conditions: 16.83¢ vs. 1.83¢, 30.54¢ vs. 3.38¢, and 21.58¢ vs. 2.24¢, respectively), and a repeated measures analysis of variance on punishment levels showed a strong main effect of the target player's behavior, $F(1, 113) = 45.19$, $p < .0001$, a marginal difference between experimental conditions, $F(2, 113) = 2.54$, $p = .083$, but no interaction between experimental conditions and the target of punishment, $F(2, 113) = 1.36$, $p = .26$.

Regarding reward, more participants provided first-order rewards to the cooperators than to the defectors in all three conditions (punisher condition: 19/41 vs. 6/41; rewarder condition: 18/37 vs. 3/37; fair-sanctioner condition: 25/38 vs. 12/38). A repeated measures analysis with weighted least squares showed that more participants rewarded cooperators than defectors, $\chi^2(1, N = 116) = 46.21$, $p < .0001$, and that the experimental conditions did differ in the number of participants who rewarded, $\chi^2(2, N = 116) = 8.23$, $p = .016$, but the interaction effect was not significant, $\chi^2(2, N = 116) = 0.56$, $p = .76$. More money was spent on first-order rewards to cooperators than was spent on defectors (for the punisher, rewarder, and fair-sanctioner conditions: 15.73¢ vs. 5.12¢, 16.08¢ vs. 4.05¢, and 28.82¢ vs. 13.68¢), and a repeated measures analysis of variance on reward levels showed a strong main effect of the target player's behavior, $F(1, 113) = 22.91$, $p < .0001$, and a significant effect of experimental condition, $F(2, 113) = 3.54$, $p = .032$, with participants in the fair-sanctioner condition spending marginally more on rewards than did participants in the punisher and reward conditions (Tukey honestly significant difference ps = .060 and .057, respectively; the latter two did not differ significantly, Tukey honestly significant difference $p = 1.00$). However, there was no significant interaction between target player's behavior and experimental con-

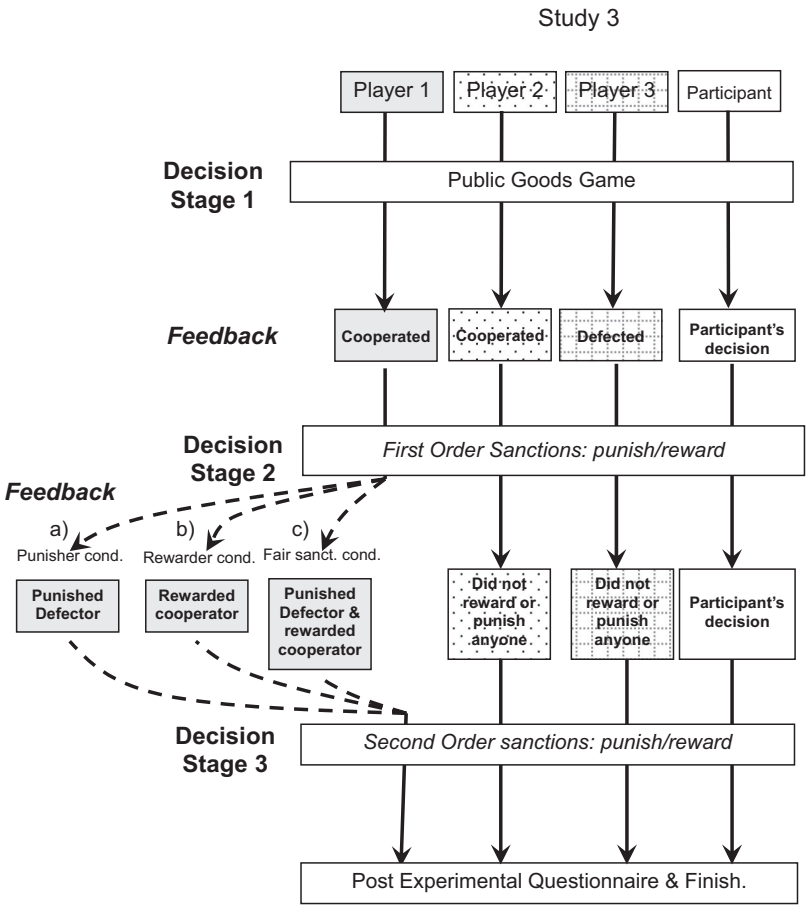

Figure 7. Flow of Study 3 from each participant's perspective in the punisher (a), rewarder (b), and fair-sanctioner (c) conditions. At each sanction stage, participants could use either punishment and/or rewards, and the three experimental conditions differed only in the particular type of sanctions used by the sanctioner (Player 1 in this figure) at the second feedback stage. As in Studies 1 and 2, the player numbers associated with each other "person" could vary.

dition, $F(2, 113) = 0.26$, $p = .771$, so we can conclude that people's differential rewarding of cooperators and defectors did not vary across the three feedback conditions before the feedback occurred.

Both Study 1 and Study 2 showed consistently that more money was spent to reward each cooperator than was spent to punish each defector. In Study 3, however, there was no statistical difference between conditions, Kruskal-Wallis test: $\chi^2(2, N = 116) = 3.76$, $p = .15$, so we aggregated the three conditions into one (total $N = 116$) and found no statistical differences between the amount spent on rewarding each cooperator and the amount spent on punishing each defector (average of 2.63¢ more on punishment; $S = 157$, $p = .46$). These results imply that having opportunities for both types of sanction at the same time makes people use them a bit more equally.

To see whether all participants reacted to cooperation and defection the same way, we classified participants according to whether or not they had themselves cooperated in the PGG. Nine of 42 (21.4%) noncooperative participants and 48 (64.9%) of 74 cooperative participants spent some money to provide first-order punishment to the defector, $\chi^2(1, N = 116) = 20.23$, $p < .0001$. Twenty-two of 42 (52.4%) noncooperative participants and 40 (54.1%) of 74 cooperative participants spent some money to provide first-order rewards to the cooperators, $\chi^2(1, N = 116) = 0.03$, $p = .86$. The two types of participants differed in the degree to which they punished defectors more than they punished cooperators (7.86¢ vs. 5.60¢ by defectors; 31.22¢ vs. 0.68¢ by cooperators, respectively; $z = 4.76$, $p < .0001$) and also differed in the degree to which they rewarded cooperators more than they rewarded defectors (17.98¢ vs. 13.81¢ by defectors, 21.35¢ vs. 4.05¢ by cooperators, respectively; $z = 2.54$, $p = .011$). Thus, cooperative participants showed clear tendencies to punish defection and reward cooperation, but noncooperative participants did not discriminate as much between cooperators and defectors in either punishment or rewards.

Second-order sanctions. At the second sanctioning stage in the punisher condition, slightly more participants provided second-order punishment to the cooperator/punisher than to the cooperator/nonsanctioner (7/41 vs. 3/41), $\chi^2(1, N = 41) = 4.43$, $p = .035$, and more money was spent on second-order punishment of the cooperator/punisher than was spent on the cooperator/nonsanctioner, though the difference was not significant (7.32¢ vs. 1.70¢, respectively; $S = 5$, $p = .13$). Furthermore, slightly fewer partic-

ipants provided second-order rewards to the cooperator/punisher than to the cooperator/nonsanctioner (9/41 vs. 12/41), and slightly less money was spent on second-order rewards to the cooperator/punisher than was spent on the cooperator/nonsanctioner (8.54¢ vs. 10.73¢, respectively; $S = 8.5$, $p = .13$), though these differences were not significant, $\chi^2(1, N = 41) = 1.88$, $p = .170$.

In the rewarder condition, an equal number of participants provided second-order punishment to the cooperator/rewarder and to the cooperator/nonsanctioner (2/37 vs. 2/37), and slightly but not significantly more money was spent to punish the latter (4.86¢ vs. 2.97¢, respectively; $S = 0.5$, $p = 1.00$). Slightly (but not significantly) more participants provided second-order rewards to the cooperator/rewarder than to the cooperator/nonsanctioner (13/37 vs. 11/37), $\chi^2(1, N = 37) = 1.03$, $p = .31$, and they spent slightly but not significantly more on second-order rewards to the former (12.97¢ vs. 8.92¢, respectively; $S = 5.5$, $p = .28$).

In the fair-sanctioner condition, more participants provided second-order punishment to the cooperator/sanctioner than to the cooperator/nonsanctioner (6/38 vs. 1/38), $\chi^2(1, N = 38) = 5.76$, $p = .016$, and more money for second-order punishment was spent on the cooperator/sanctioner than was spent on the cooperator/nonsanctioner (4.74¢ vs. 0.53¢, respectively; $S = 8.5$, $p = .13$), but this latter difference did not reach significance. On the other hand, slightly more participants provided second-order rewards to the cooperator/sanctioner than to the cooperator/nonsanctioner (15/38 vs. 12/38), $\chi^2(1, N = 38) = 1.89$, $p = .17$, and more money was spent to reward the cooperator/sanctioner than was spent on the cooperator/nonsanctioner, and this latter effect was nonsignificant (18.68¢ vs. 13.95¢, respectively; $S = 9.5$, $p = .063$).

To test for the total advantage of the cooperator/sanctioner over the cooperator/nonsanctioner in each condition, we calculated the amount of punishment toward the nonsanctioner minus the punishment toward the sanctioner, and also the amount of rewards toward the sanctioner minus the rewards toward the nonsanctioner, and then added these two together. If the cooperator/sanctioner is advantaged relative to the cooperator/nonsanctioner, this number will be positive. Using this measure, the cooperator/punisher in the punisher condition was worse off relative to the cooperator/nonsanctioner (−7.80¢; $S = 12.5$, $p = .047$), the cooperator/rewarder in the rewarder condition was slightly but not significantly better off relative to the cooperator/nonsanctioner (+5.95¢; $S = 9$, $p = .14$), and the fair sanctioner (who cooperates, punishes the defector, and rewards the cooperator) was just as well off as the cooperator/nonsanctioner (+0.53¢; $S = 0$, $p = 1.00$). The three experimental conditions were significantly different, Kruskal-Wallis test: $\chi^2(2, N = 116) = 6.36$, $p = .042$, with the sanctioner being worse off relative to the nonsanctioner in the punishment condition than the sanctioner was relative to the nonsanctioner in the reward condition ($z = 2.68$, $p = .007$), with the fair sanctioner in between those two and not significantly different from either ($z = 1.46$ and 0.92, ps $= .14$ and .36, respectively). This supports the conclusions of Studies 1 and 2 that punishing brings negative reactions from others, rewarding brings positive reactions, and these effects cancel each other out when one does both.

Postexperimental questionnaires. Once again, cooperators were evaluated more favorably than defectors were in all conditions (average differences in the punisher, rewarder, and fair-sanctioner conditions: +2.89, +2.80, and +2.50, respectively; Ss $= 389$, 333, and 348.5, respectively, all ps $< .0001$). Both cooperative and noncooperative participants evaluated the cooperators more positively than they did the defectors in all conditions (average differences by cooperative and noncooperative participants in the punisher condition: +3.36 [$n = 24$] vs. +2.24 [$n = 17$]; rewarder condition: +3.24 [$n = 25$] vs. +1.88 [$n = 12$]; and fair-sanctioner condition: +2.77 [$n = 25$] vs. +1.97 [n=13]). Cooperative subjects showed greater discrimination against defectors, $F(1, 110) = 8.32$, $p = .0047$, and this did not differ between conditions or interact with experimental condition (both Fs < 1).

In the punisher condition, the cooperator/punishers were evaluated slightly less favorably than were the cooperator/nonsanctioners (average differences in ratings: −0.09), but this difference was not significant ($S = 20$, $p = .74$). Cooperative ($n = 24$) and noncooperative ($n = 17$) participants both evaluated the cooperator/punisher slightly less favorably than they did the cooperator/nonsanctioner, and their differences in ratings were not significantly different (−0.03 vs. −0.18, respectively; $z = 0.48$, $p = .63$). In the rewarder condition, the cooperator/rewarders were evaluated significantly more favorably than were the cooperator/nonsanctioners (average differences in ratings: +0.36; $S = 135$, $p = .009$). The cooperator/rewarder was rated more favorably than was the cooperator/nonsanctioner by cooperative ($n = 25$) and noncooperative ($n = 12$) participants alike (average difference in ratings: +0.31 and + 0.46; Ss $= 58.5$ and 18, ps $= .038$ and .12, respectively), and although this effect was only significant for the cooperative participants, the difference in ratings between cooperative and noncooperative participants was not significantly different ($z = 0.18$, $p = .86$). In the fair-sanctioner condition, the cooperator/sanctioners received approximately the same average ratings as did the cooperator/nonsanctioners (average difference in ratings: +0.06; $S = 37$, $p = .52$). Cooperative participants ($n = 25$) evaluated the cooperator/sanctioners slightly more positively than they did the cooperator/nonsanctioner (+0.37; $S = 50.5$, $p = .078$), but noncooperative participants ($n = 13$) evaluated them slightly more negatively (−0.53; $S = 14$, $p = .30$), and this difference in ratings between cooperative and noncooperative participants was nonsignificant ($z = 1.87$, $p = .062$). This supports the notion that people readily endorse positive sanctions but not negative sanctions, and only cooperative participants appreciate the use of both types of sanction.

Summary of Study 3

The presence of both punishment and rewards makes situations more complex. As in Studies 1 and 2, participants punished defectors more and rewarded them less than they did cooperators, and although rewards were used by cooperative and uncooperative participants, punishment tended to be used more by the former. Unlike Studies 1 and 2, however, there was no clear preference for using rewards rather than punishment at the first-order level. With second-order sanctions, the pattern of rewarding rewarders but not punishing nonpunishers was visible, but it was not significant in Study 3 when both types of sanction were available (although the trend toward rewarding sanctioners did approach significance in the fair-sanctioner condition, and participants did rate rewarders positively in the postexperimental questionnaires). One thing is clear and that is despite the fact that nonsanctioners provided no positive or negative sanctions (which makes it more likely that not punishing is due to selfishness), they were still not punished more

than were punishers. In other words, there was no second-order punishment of nonpunishers, and if anything, punishers were disadvantaged relative to nonpunishers. Comparing the punisher and rewarder conditions shows that those who reward end up better off than do those who punish.

General Discussion

Rewards Versus Punishment

Several findings were consistent across all three studies. Those who contributed to public goods received more reward and less punishment than did those who did not contribute. Cooperation was rewarded and viewed positively by cooperators and noncooperators alike, but only cooperators tended to punish defection. At the level of second-order sanctions, participants consistently showed positive responses toward those who used rewards but not toward those who used punishment (although this effect was not as strong when people could use both types of sanctions simultaneously), and rewarders received more favorable outcomes than did punishers in all three studies. Nonpunishers did not receive more punishment than those who did punish, and in fact, the former may have received less punishment, even in Study 3 when the nonpunisher was clearly unwilling to support cooperation with either positive or negative sanctions (making his/her motives for not punishing seem less like mercy and more like free riding on others' sanctions). This all suggests that people will much more readily support positive sanctions than they will support negative sanctions. These results are consistent with past research that has shown that cooperation is higher when incentives are framed as rewards rather than as punishments (Komorita, 1987; Komorita & Barth, 1985) and that reinforcement-based learning is often more effective and desirable than is punishment-based learning (Skinner, 1971).

Theoreticians have argued that higher order punishment can stabilize lower order punishment because the cost of providing second-order sanctions is low once cooperation is common (e.g., Boyd et al., 2003; Sober & Wilson, 1999). People clearly possess a psychology that makes them want to punish noncooperators, even in one-shot interactions (e.g., Fehr & Gächter, 2002), and the present studies support this. However, our results also suggest that people—at least participants in our studies—do not readily engage in second-order punishment of nonpunishers. This supports other research from Kiyonari et al. (2008) in Japanese participants and Barclay (2006) in multiple-round interactions, although the latter did not have a separate stage to explicitly test for second-order sanctions. With this apparent lack of support for punishing, it becomes difficult for the norm of punishment to become established and makes second-order punishment unlikely to be a major force in stabilizing cooperation. Conversely, our results show that higher order rewards are readily given: Our participants selectively rewarded cooperators and rewarders who rewarded cooperators. In short, despite the power of punishment to change behavior in the short term, punishment cannot easily sustain itself through higher order punishment, yet rewarding may be able to sustain itself through higher order rewarding, and people's actual behavior in these studies supports these intuitions from evolutionary biology.

Several studies have shown that rewards can sustain cooperation in repeated interactions (e.g., McCusker & Carnevale, 1995; Milinski et al., 2002; Rockenbach & Milinski, 2006; Sefton, Shupp, & Walker, 2007), especially when the amount received by recipients is greater than the cost to the rewarder is (Vyrastekova & van Soest, 2008), but there are some theoretical problems with the use of rewards (of either first or second order) to stabilize cooperation. Rewarding becomes more costly as the number of cooperators increases because one has to reward more people (Oliver, 1980), and the increased use of rewards can lead to "inflation," whereby each additional reward does not help the recipient as much (K. Panchanathan, personal communication, September 30, 2005), which undermines our assumption that rewards benefit the recipient more than they cost the donor. However, these problems are reduced if the rewards are part of an ongoing system of indirect reciprocity or generalized exchange whereby nonrewarders simply get excluded from this already present system, just as any other type of noncooperator would. These systems of indirect reciprocity are stable in at least some forms (Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2003, 2004), and the present studies suggest that people do behave as predicted by these models by rewarding those who reward others. If indeed rewards are seen as part of a system of indirect reciprocity, then we would expect that those who provide second-order rewards would receive more help than those who do not, just as occurs with other forms of generosity under indirect reciprocity (e.g., Wedekind & Milinski, 2000) and as occurred with first-order rewarders in the present studies. Furthermore, those who use rewards might be chosen more often as social partners, such that rewards are used as a form of competitive altruism to influence partner choice (e.g., Barclay & Willer, 2007).

Proximate Psychological Mechanisms

What psychological mechanisms are involved when receiving rewards or punishment? If people respond to prosocial acts with reciprocation, whether due to increased empathy or closeness or conscious desires for reciprocity (Singer et al., 2006; Tooby & Cosmides, 1996), then second-order rewarding will naturally result. If punishment is indeed related to anger (Eisenberger et al., 2004; Fehr & Gächter, 2002), then it seems unlikely that nonpunishers should incur the wrath of others unless such nonpunishment explicitly hurts a punisher or weakens his/her position or if punishment is required from everyone in order to be effective, as is the case with countries who do not support trade embargoes. Thus, it is easy to see how the psychological mechanisms that respond to first-order reward could naturally produce second-order reward, but this is not always the case for punishment. Although extrinsic rewards can undermine intrinsic motivation (e.g., Deci, Koestner, & Ryan, 1999), which could reduce cooperation after rewards are removed, this is unproblematic so long as the system of indirect reciprocity endures, and maintaining behavior by punishment also undermines intrinsic motivation (Taylor, 1976; Yamagishi, 1988), possibly more so. Verbal rewards can enhance intrinsic motivation (Deci et al., 1999), so we suggest that verbal and other social rewards (e.g., smiles) could be the immediate reinforcers (or *secondary reinforcers* in the parlance of learning theory) for cooperation that eventually cause internalization of the norm. Responding to verbal rewards would be adaptive if such rewards are indicators of more tangible future rewards, such as increased social closeness or coalitional support, but this remains to be tested.

Although our findings suggest that people will more readily support the use of rewards than of punishment in the maintenance of cooperation, and we argue that this says something about humans' evolved psychology as produced by natural selection acting at the level of the individual (as opposed to at the group level), there are some possibilities that should be addressed. One possibility is that people used first- and second-order reward more than they did punishment because rewarding is group beneficial: Rewards increased the average payoff of the entire group, whereas punishment reduced the average payoff of the group. We readily admit that this might explain the tendency to use rewards more than punishment at the first- and second-order levels. However, this point does not predict the discriminative use of rewards—if people use rewards simply because they are group beneficial, there is less reason for them to withhold those rewards from certain individuals. Also, when participants had both sanctions available (Study 3), they did not show a preference for using rewards, suggesting that a simple preference for group-beneficial outcomes cannot explain our result that rewarders were preferred to nonrewarders, but punishers were not preferred to nonpunishers. A similar objection is that participants construed the rewards as a second and third PGG rather than as rewards for cooperation, and much research has shown that many people will cooperate in one-shot PGGs (e.g., Davis & Holt, 1993; Fehr & Gächter, 2002; Ledyard, 1995). The fact that both defectors and cooperators rewarded the first-stage cooperators is consistent with this interpretation, but two facts speak against this interpretation. First, we explicitly used the words *punishment* and *rewards* in the instructions rather than more neutral terms because we wanted to reduce ambiguity in the interpretation.[2] Second, people showed discriminative rewarding: They rewarded cooperators more than they did noncooperators and rewarders more than they did nonrewarders. This suggests that participants did indeed conceptualize the rewards as rewards rather than as second and third PGGs independent of the first.

What Does Support Punishment?

If first-order punishment so clearly exists, but second-order punishment and reward do not seem to sustain it, then what *does* maintain punishment? Some have argued that natural selection at the level of the group will cause groups with punishers to do better than groups without punishment will do (e.g., Boyd et al., 2003), but evolutionary scientists are hesitant to accept arguments based on such *group selection* (e.g., Burnham & Johnson, 2005; Dawkins, 1976). There are some individual-level benefits that punishers could accrue. If punishment of noncooperators is a demonstration of one's dislike of unfairness, then these justified punishers may be trusted more than nonpunishers are, and Barclay (2006) provided evidence that this is the case among participants who have had sufficient exposure to noncooperators and punishment. Trusting someone is different from rewarding them, in that one has a selfish incentive to trust a trustworthy person in the hopes of some return, and it is important to note that Barclay found evidence that people *trusted* punishers but no evidence of people being *nicer* to punishers (only the latter was tested in the present studies). This could result in punishers being chosen as cooperative partners or as leaders more than nonpunishers are, but this remains to be tested. Demonstrating one's dislike of unfairness or anger toward free riders may have other benefits also: People may be less willing to defect on those who are known for punishing. This is predicted by some mathematical models (Brandt, Hauert, & Sigmund, 2003; Hauert, Haiden, & Sigmund, 2004; Sigmund, Hauert, & Nowak, 2001) and has recently been verified in humans (Barclay, 2008). By demonstrating that one experiences anger toward defectors and that one will irrationally punish them even in one-shot situations, punishers demonstrate that it is not in others' best interests to defect on the punisher. Although each single instance of punishment may not be in one's best interest, the presence of these irrational emotions commits us to carrying out the threat of punishment and, as such, is beneficial in the long run because it makes the threat of punishment credible (Frank, 1988).

Yamagishi and Takahashi (1994) provided another possible explanation for the existence of punishment: behavioral linkage between cooperation and first-order punishment, such that only cooperators perform punishment, whereas free riders do not. Indeed, our study shows some evidence for behavioral linkage: 63% (82/131) of all cooperative participations but only 22% (16/73) of noncooperative participants punished defectors. Although not complete, linkage is clearly present and could be caused by a mechanism, like fairness concerns, simultaneously causing cooperation and dislike (and hence punishment) of noncooperation. Linkage may not completely solve the problem, however, because it is unclear whether such linkage could evolve or be stable in larger cooperative groups where nonpunishers are likely to share an interacting group with punishers and personally benefit from the cooperation enforced by the latter. Interestingly, we found no linkage between cooperation and rewards: Cooperative and noncooperative participants alike provided selective rewards. This suggests that the two types of sanctions (punishment and reward) could arise from very different evolutionary processes and psychological mechanisms, even if they are identical from a rational choice perspective.

We should also discriminate between institutional sanctioning systems and voluntary peer-to-peer sanctioning and note that our focus in this article is mostly on the latter. The direct personal benefits of punishing can sometimes outweigh the personal costs of doing so (West, Griffin, & Gardner, 2007), but when these benefits are small relative to the cost of punishment and retaliation, peer-to-peer punishment may not play a big role in large-scale cooperation. In modern societies, punishment systems are usually provided by central authorities, such as governments and leaders, and people pay for punishment systems (e.g., police) through taxes or fines. Leaders may benefit from punishing noncooperators, even if they do not personally benefit from the increased cooperation, if their reputation as a leader increases from their ability to maintain cooperation or if it is an expected part of their role. This connection between punishment and leadership or social status has received little investigation (but see Barclay, 2005) and deserves further theoretical and empirical work.

[2] These terms are unlikely to have caused an unwillingness to punish nonsanctioners because participants were quite willing to punish noncooperators at the first stage.

Conclusions

In summary, we show that people readily provide first-order punishment of noncooperators but not second-order punishment of nonpunishers, suggesting that people do not behave as predicted by any theories of behavior that rely on second-order punishment to sustain cooperation (e.g., Axelrod, 1986; Boyd & Richerson, 1992; Henrich, 2004; Henrich & Boyd, 2001). In contrast, people readily provide first-order rewards to cooperators *and* second-order rewards toward those who reward cooperators, as predicted by theories of indirect reciprocity. As such, we argue that cooperation is more likely to be maintained by systems of reward than by systems of punishment because the former systems are more easily supported by higher order sanctions and are thus more likely to persist.

References

Allport, G. W. (1955). *Becoming: Basic considerations for a psychology of personality.* New Haven, CT: Yale University Press.

Axelrod, R. (1984). *The evolution of cooperation.* New York: Basic Books.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review, 80,* 1095–1111.

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons." *Evolution & Human Behavior, 25,* 209–220.

Barclay, P. (2005). *Reputational benefits of altruism and altruistic punishment*. Unpublished doctoral dissertation, McMaster University, Hamilton, Ontario, Canada.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27,* 325–344.

Barclay, P. (2008). *"Don't mess with the enforcer": Deterrence as an individual-level benefit for punishing free-riders.* Manuscript submitted for publication.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society of London B: Biological Sciences, 274,* 749–753.

Batson, C. D., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K. (1997). Is empathy-induced helping due to self-other merging? *Journal of Personality and Social Psychology, 73,* 495–509.

Baumeister, R. F. (1998). The self. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 680–740). New York: McGraw-Hill.

Bonacich, P., Shure, G. H., Kahan, J. P., & Meeker, R. J. (1976). Cooperation and group size in the N-person prisoners' dilemma. *Journal of Conflict Resolution, 20,* 687–706.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, USA, 100,* 3531–3535.

Boyd, R., & Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology, 132,* 337–356.

Boyd, R., & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13,* 171–195.

Brandt, H., Hauert, C., & Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London B: Biological Sciences, 270,* 1099–1104.

Brandt, H., Hauert, C., & Sigmund, K. (2006). Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences, USA, 103,* 495–497.

Burnham, T., & Johnson, D. (2005). The biological and evolutionary logic of human cooperation. *Analyse & Kritik, 27,* 113–135.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics, 9,* 265–279.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics.* Princeton, NJ: Princeton University Press.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31,* 169–193.

Dawkins, R. (1976). *The selfish gene.* Oxford, UK: Oxford University Press.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on internal motivation. *Psychological Bulletin, 125,* 627–668.

de Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004, August 27). The neural basis of altruistic punishment. *Science, 305,* 1254–1258.

Eisenberger, R., Lynch, P., Aselage, J., & Rohdieck, S. (2004). Who takes the most revenge? Individual differences in negative reciprocity norm endorsement. *Personality and Social Psychology Bulletin, 30,* 787–799.

Fehr, E., & Fischbacher, U. (2003, October 23). The nature of human altruism. *Nature, 425,* 785–791.

Fehr, E., & Fischbacher, U. (2004). Third party punishment and social norms. *Evolution and Human Behavior, 25,* 63–87.

Fehr, E., & Gächter, S. (2002, January 10). Altruistic punishment in humans. *Nature, 415,* 137–140.

Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences, USA, 102,* 7047–7049.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions.* New York: W. W. Norton.

Gigerenzer, G. (2001). Decision making: Nonrational theories. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 5, pp. 3304–3309). Oxford, UK: Elsevier.

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology, 69,* 339–348.

Hardin, G. (1968, December 13). The tragedy of the commons. *Science, 162,* 1243–1248.

Hardin, R. (1971). Collective action as an agreeable *n*-prisoners' dilemma. *Behavioral Science, 16,* 472–481.

Hardy, C., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin, 32,* 1402–1413.

Hauert, C., Haiden, N., & Sigmund, K. (2004). The dynamics of public goods. *Discrete and Continuous Dynamical Systems B, 4,* 575–587.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization, 53,* 3–35.

Henrich, J., & Boyd, R. (2001). Why people punish defectors—Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology, 208,* 79–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28,* 795–855.

Jones, S. C. (1973). Self- and interpersonal evaluations: Esteem theories versus consistency theories. *Psychological Bulletin, 79,* 185–199.

Kiyonari, T., van Veelen, M., & Yamagishi, T. (2008). *Can second-order punishment solve the puzzle of human cooperation in one-shot games?* Manuscript submitted for publication.

Komorita, S. S. (1987). Cooperative choice in decomposed social dilemmas. *Personality & Social Psychology Bulletin, 13,* 53–63.

Komorita, S. S., & Barth, J. M. (1985). Components of reward in social dilemmas. *Journal of Personality and Social Psychology, 48,* 364–373.

Komorita, S. S., & Parks, C. D. (1995). Interpersonal relations: Mixed-motive interaction. *Annual Review of Psychology, 46,* 183–207.

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 111–194). Princeton, NJ: Princeton University Press.

Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London B: Biological Sciences, 268,* 745–753.

McCusker, C., & Carnevale, P. J. (1995). Framing in resource dilemmas: Loss aversion and the moderating effects of sanctions. *Organizational Behavior and Human Decision Processes, 61,* 190–201.

Messick, D. M., & Brewer, M. B. (1983). Solving social dilemmas: A review. In L. Wheeler & P. Shaver (Eds.), *Review of personality and social psychology: Vol. 4* (pp. 11–44). Beverley Hills, CA: Sage Publications.

Milinski, M., Semmann, D., & Krambeck, H. J. (2002, January 24). Reputation helps solve the 'tragedy of the commons.' *Nature, 415,* 424–426.

Milinski, M., Semmann, D., Krambeck, H. J., & Marotzke, J. (2006). Stabilizing the earth's climate is not a losing game: Supporting evidence from public goods experiments. *Proceedings of the National Academy of Sciences, USA, 103,* 3994–3998.

Nowak, M., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology, 194,* 561–574.

Oliver, P. (1980). Rewards and punishment as selective incentives for collective action: Theoretical investigations. *American Journal of Sociology, 85,* 1356–1375.

Olson, M., Jr. (1965). *The logic of collective action.* Cambridge, MA: Harvard University Press.

Ostrom, E. (1990). *Governing the commons.* New York: Cambridge University Press.

Ostrom, E. J., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review, 86,* 404–417.

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: The importance of standing for the evolution of reciprocity. *Journal of Theoretical Biology, 224,* 115–126.

Panchanathan, K., & Boyd, R. (2004, November 25). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432,* 499–502.

Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology, 28,* 363–392.

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society of London B: Biological Sciences, 265,* 427–431.

Rockenbach, B., & Milinski, M. (2006, December 7). The efficient interaction of indirect reciprocity and costly punishment. *Nature, 444,* 718–723.

Samuelson, P. A. (1954). The pure theory of public expenditure. *Review of Economics and Statistics, 36,* 387–390.

Sefton, M., Shupp, R., & Walker, J. (2007). The effects of rewards and sanctions in provision of public goods. *Economic Inquiry, 45,* 670–690.

Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology, 80,* 325–339.

Sigmund, K., Hauert, C., & Nowak, M. A. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences, USA, 98,* 10757–10762.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006, January 26). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439,* 466–469.

Skinner, B. F. (1971). *Beyond freedom and dignity.* New York: Knopf.

Sober, E., & Wilson, D. S. (1999). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Taylor, M. (1976). *Anarchy and cooperation.* New York: Wiley.

Tooby, J., & Cosmides, L. (1996). Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. *Proceedings of the British Academy, 88,* 119–143.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46,* 35–57.

U.S. Department of State International Information Programs (n.d.) *Cuban Liberty and Democratic Solidarity (Libertad) Act of 1996.* Retrieved December 13, 2007, from http://usinfo.state.gov/regional/ar/us-cuba/libertad.htm

Vogel, G. (2004, February 20). The evolution of the golden rule. *Science, 303,* 1128–1131.

Vyrastekova, J., & van Soest, D. (2008). On the (in)effectiveness of rewards in sustaining cooperation. *Experimental Economics, 11,* 53–65.

Wedekind, C., & Milinski, M. (2000, May 5). Cooperation through image scoring in humans. *Science, 288,* 850–852.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity, and group selection. *Journal of Evolutionary Biology, 20,* 415–432.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51,* 110–116.

Yamagishi, T. (1988). Exit from the group as an individualistic solution to the free rider problem in the United States and Japan. *Journal of Experimental Social Psychology, 24,* 530–542.

Yamagishi, T., & Takahashi, N. (1994). Evolution of norms without metanorms. In U. Schulz, W. Albers, & U. Mueller (Eds.), *Social dilemmas and cooperation* (pp. 311–326). Berlin, Germany: Springer-Verlag.

Received August 23, 2007
Revision received January 5, 2008
Accepted January 7, 2008 ■