False friends are worse than bitter enemies: “Altruistic” punishment of in-group members

Mizuho Shinada, Toshio Yamagishi*, Yu Ohmura

Graduate School of Letters, Hokkaido University, N10 W7 Kita-ku, Sapporo 060-0810, Japan

Initial receipt 9 April 2004; final revision received 3 August 2004

Abstract

One of the most critical features of human society is the pervasiveness of cooperation in social and economic exchanges. Moreover, social scientists have found overwhelming evidence that such cooperative behavior is likely to be directed toward in-group members. We propose that the group-based nature of cooperation includes punishment behavior. Punishment behavior is used to maintain cooperation within systems of social exchange and, thus, is directed toward members of an exchange system. Because social exchanges often take place within groups, we predict that punishment behavior is used to maintain cooperation in the punisher’s group. Specifically, punishment behavior is directed toward in-group members who are found to be noncooperators. To examine this, we conducted a gift-giving game experiment with third-party punishment. The results of the experiment ($N = 90$) support the following hypothesis: Participants who are cooperative in a gift-giving game punish noncooperative in-group members more severely than they punish noncooperative out-group members. © 2004 Elsevier Inc. All rights reserved.

Keywords: Social exchange; Third-party punishment; Gift-giving game; Group-based cooperation; Second-order cooperation

* Corresponding author. Tel.: +81 11 706 4157; fax: +81 11 706 03066.
E-mail address: toshio@let.hokudai.ac.jp (T. Yamagishi).
1. Introduction

One of the most critical features of human society is the pervasiveness of cooperation in social and economic exchanges. From cooperative hunting and the sharing of meat in hunter-gatherer societies (Gurven, Allen-Arave, Hill, & Hurtado, 2000; Hawkes, 1993; Kaplan & Hill, 1985) to market trades in industrial societies, people engage in cooperative activities and social exchanges in both the public and private domains of their lives. One of the most essential adaptive tasks for humans is to achieve mutual cooperation in social exchanges, particularly in situations where material incentives to cheat exchange partners are rampant (Cosmides, 1989; Cosmides & Tooby, 1989, 1992). Why do self-interested individuals cooperate even when they can obtain better outcomes by cheating others? This has been a controversial issue in the behavioral sciences for at least the past two decades. Many answers have been proposed. Most researchers agree with evolutionary theorists that there are two important principles that help us understand this phenomenon: kin selection (Hamilton, 1964a, 1964b) and reciprocal altruism (Axelrod & Hamilton, 1981; Trivers, 1971). Kin selection explains cooperative behavior among genetically related actors even when that behavior is costly. Reciprocal altruism helps us understand how mutual cooperation in bilaterally repeated interactions among nonkin members occurs. Taken together, kin selection and reciprocal altruism explain a considerable degree of cooperative behavior in human societies.

While kin selection and reciprocal altruism address costly cooperation among kin or among nonkin members in repeated interaction, respectively, costly cooperation in social exchanges among genetically unrelated actors beyond repeated dyadic interactions occurs often in human societies. The extensive practice of costly cooperation among genetically unrelated actors beyond repeated dyadic relations is a challenge that drives theorists to search for a plausible explanation. In response to this challenge, several models that explain the evolution of cooperation in sizeable groups have been proposed. Examples of such models include indirect reciprocity (Alexander, 1987; Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998), group selection (Gintis, 2000; Gintis, Bowles, Boyd, & Fehr, 2003), and cultural group selection (Henrich & Boyd, 2001). A common feature in these models is that the cooperation they examine takes place among a limited set of actors. According to the indirect reciprocity model, for example, cooperation takes place among a group of people in which reputation information can be shared. According to the group selection and cultural group selection models, cooperation takes place within demarcated groups. Recently, Axelrod and Hammond (2003) proposed a theoretical model that illustrates the possibility that in-group favoritism (ethnocentric behavior) can evolve even among actors who are endowed with minimal memory and cognitive ability.

Empirical findings are consistent with the model’s implications. A number of experimental studies by social psychologists show that people cooperate with their own group members more than with strangers. For example, Dawes, van de Kragt, and Orbell (1988) find that the provision of public goods is considerably higher when the benefits of cooperation go to in-group members than when they go to out-group members. In their experiment, the positive effect of group identity on cooperative behavior was observed even when the group identity
was induced arbitrarily by the experimenter. Likewise, researchers using the “minimal group paradigm” (e.g., Tajfel & Turner, 1979) demonstrate that in-group favoritism—that is, preferential treatment of in-group members over out-group members in reward allocation—occurs even when participants are arbitrarily divided into two groups according to a trivial trait (e.g., Rabbie, Schot, & Visser, 1989; Tajfel, Billig, Bundy, & Flament, 1971; Yamagishi, Jin, & Kiyonari, 1999).

As briefly shown above, experimental evidence illustrates that cooperative behavior is more often directed toward in-group members than toward out-group members. Such group-based cooperation, or “in-group cooperation,” occurs even in one-shot experimental games in which cooperative behavior cannot generate future benefits to the players. Costly cooperation in one-shot games is considered an altruistic behavior because the player who cooperates cannot expect future benefits of his or her cooperative behavior. Such “altruistic” behavior could possibly be produced by a set of psychological mechanisms that are adaptive in everyday life, especially in the environment of evolutionary adaptedness (EEA). In other words, altruistic cooperation in one-shot games is likely only to the degree that it is based on psychological mechanisms that were adaptive in the EEA. The fact that altruistic cooperation occurs mostly with in-group members suggests that it is a product of psychological mechanisms that are designed to promote mutual cooperation (or indirect reciprocity) within-group boundaries (Yamagishi et al., 1999).

The goal of this study is to demonstrate that the group-based nature of cooperation is extended to punishment behavior or what is referred to as second-order cooperation. The punishment of noncooperators constitutes a so-called second-order social dilemma (e.g., Yamagishi, 1986a) because it forces free riders to change their behavior and cooperate so that all members of the group benefit. Engaging in punishment behavior provides benefits to all members of a group, but engaging in it can be costly to the punisher (Fehr & Gächter, 2002; Oliver, 1980; Yamagishi, 1986a). Engaging in punishment behavior is costly, and the provision of punishment is itself a public good. If the goal of punishment is to improve the behavior of potential free riders, then punishment should be directed to potential free riders in the punisher’s own group, where the benefits of cooperation are shared among group members, including the punisher. We argue that engaging in punishment behavior to induce cooperation is more likely to be directed toward in-group members than toward out-group members in the same manner that cooperative behavior is more likely to be directed toward in-group members than toward out-group members.

To empirically demonstrate the prediction that punishment is more often directed toward in-group than out-group members, we conducted a third-party punishment game (TPG; Fehr & Fischbacher, 2004). Experimental economists invented an ingenious game to test whether humans have a tendency to punish individuals who violate prosocial norms, even when there is a substantial cost to them and even when the norm violations do not directly harm the punisher. A TPG is played among three actors—one game player and a third party—as a one-shot game. Two actors play a dictator game or a prisoner’s dilemma game. When the TPG involves a dictator game, it is played by a dictator, a recipient, and a third party. The “dictator,” who is endowed with a fixed amount of money, decides to transfer some, none, or all of his endowment to a passive “recipient,” who is given no endowment. The third party,
who is also endowed with money, is informed of the amount that the dictator transferred to
the recipient and is provided with an opportunity to punish the dictator (i.e., to reduce the
dictator’s payoff) by spending some or all the endowment for punishment. When the TPG
involves a PD game rather than a dictator game, it is played by two PD players and a third
party. The PD players first play a one-shot PD game. The third party is informed of the two
players’ behavior in the PD game and then decides whether to punish the PD players.
Punishment in the TPG, whether it involves a dictator or a PD game, is not only costly for the
sanctioned dictator or PD player but also for the third party who punishes. A self-interested
third party is thus expected not to punish. We call punishment in the TPG altruistic
punishment because it never brings the punisher benefits in the form of a positive reputation
or in the form of an improved level of cooperation of the sanctioned player. We notice that the
use of the term altruistic here may not be completely appropriate because punishment in the
TPG does not produce any positive effect, such as altering the behavior of the norm violator.
That is, punishment in the TPG cannot benefit other players. Despite this definitional
problem, we speculate that the same adaptive psychological mechanisms that produce
altruistic cooperation toward in-group members produce punishment of in-group norm
violators. Both altruistic cooperation and altruistic punishment, we speculate, are products of

Contrary to the prediction that the third party will not punish in the TPG (based on the
assumption of rational actors), results of the TPG experiments consistently indicate that a
substantial proportion of participants incur costs to themselves to punish a norm violator. The
norm violators who are punished are the dictators who transfer their endowment selfishly or
the PD players who defect in the prisoner’s dilemma when their partner cooperates. We use a
TPG to examine if altruistic punishment, as costly cooperation in the second-order social
dilemma, is more often directed toward in-group members than toward out-group members.
In our version of the TPG, we used a gift-giving game instead of a dictator or PD game. A
gift-giving game is similar to a dictator game. One player transfers a portion of his or her
endowment to another player. The gift-giving game differs from a dictator game in that the
transferred money is doubled by the experimenter before it reaches the recipient. In other
words, the transfer of money in a dictator game is not value adding, whereas the transfer of
money in a gift-giving game is. We believe that the latter situation is more representative of
social exchange than is the former situation. In social exchange settings, we exchange
resources whereby the exchange benefits both actors. Thus, value is added to our resources
through the exchange.

Another important difference between the standard TPG and our TPG is that our game is
played among eight players rather than three. The eight players are divided into two groups
of four players each. Participants are a member of one of the two four-person groups: The group
that they are a member of is the in-group, and the group they are not a member of is the out-
group. A gift-giving game is played within each group. In each group, three of the four
players are endowed with money. They decide how much money to give to another player in
a unidirectional triangle (A→B→C→A). The remaining member of the four-person group
is assigned the role of an observer, who observes the decisions made by players who are
giving gifts. After observing these decisions, observers decide how much of their endowment
to spend in punishing each of the players. The observer can punish in-group and out-group players. The details of the procedure will be presented later.

We predict that participants in this experiment will punish noncooperative players (i.e., the gift-giving players who do not give to another player) more severely when those players are in-group members than when they are out-group members. This hypothesis of in-group punishment is based on the assumption that punishment is a form of second-order cooperation. There is some evidence that first-order cooperation correlates with second-order cooperation (Fehr & Gachter, 2002; Price, Cosmides, & Tooby, 2002). Furthermore, we predict that those who do not cooperate in the gift-giving game (first-order noncooperators) are not likely to engage in the second-order cooperation of punishment. As a result, we limit our prediction of in-group punishment to those who cooperate in the gift-giving game. We have no prediction concerning noncooperators in the gift-giving game—if they punish at all for some reason, whether they punish in-group members more than out-group members. As will be shown later, the results of our experiment support our hypothesis: Cooperators in the gift-giving game punish in-group “cheaters” more severely than they do out-group cheaters.

2. The experiment

2.1. Procedure

The experiment was conducted in a laboratory consisting of 16 small rooms. Upon arrival to the laboratory complex, each participant was immediately led to one of the rooms to minimize the chances that participants might meet one another. The experiment was conducted on a computer located in each room. As stated earlier, participants were led to believe that there were eight actors necessary for the experiment, and they were divided into two 4-person groups. However, only one of the eight participants was a real participant. The other seven “participants” did not exist. The participant was led to believe, in the instructions, that there are seven other people in the laboratory complex. Several participants participated in the experiment at the same time in different rooms. Because the rooms were not completely sound proof, participants felt that other participants were present. Such a situation reduced any suspicion among the participants that they were the only real participants. When there were only a few participants, the experimenter would walk into empty rooms nearby and begin giving instructions loudly as if there were other participants in those rooms. None of the participants expressed their suspicion concerning the other participants in the postexperimental questionnaire.

2.1.1. Gift-giving game

The experiment begins with instructions shown on the participant’s computer monitor. After the participants are provided with a show-up fee of 300 yen (that they can later use to punish other participants), they are told that two 4-person groups have been formed. They are informed that one of the groups consists of students who attend the same gakubu (academic unit, equivalent of school or college in American universities) that they attend and that the
other group consists of students who attend another gakubu. Next, participants are told that they will play a gift-giving game repeatedly in each group. In each round of the game, three of the four members of each group are randomly assigned the role of “traders” and the remaining member is assigned the role of an “observer.” Traders are endowed with 200 yen (about US$2.00) each and decide how much of that money to give to another trader of their group. Each trader gives his or her money to one other trader and is then given money by the third trader in a unidirectional triangle. Thus, the within-group gift giving constitutes a chain-generalized exchange system (Ekeh, 1974). Note that no direct reciprocity is possible in this system. The amount of money that a member gives is doubled by the experimenter, and thus, the recipient of the gift receives twice as much as the gift giver gives away. If everyone gives the full endowment of 200 yen, each receives 400 yen. In this sense, the generalized exchange system is a public good. Giving money is cooperation for the maintenance of the generalized exchange system, and not giving money is free riding on the generalized exchange system. At the end of each round, members are informed of the amount given by another trader of their own group. Participants are told that there will be several rounds of trading, but they are not told exactly how many rounds there will be.

2.1.2. Observer’s role

In each round, one of the four members of each group is assigned the role of an observer. Thus, there are two observers in each round, one from each group. The observer does not participate in the gift-giving system. With a probability unknown to the participants, each observer is provided with an opportunity to observe how much each of the other six traders (including traders from their own group and traders from the other group) gave in that round, and then to punish any one of them. When this opportunity is provided, the observer decides how much they pay (from their own income) to punish others. How the punishment is executed will be presented later. It is important, however, to note that traders are not informed until the end of all rounds whether they received a punishment. Thus, punishments in this game are totally useless as a means to affect the behavior of other participants. In other words, no long-term benefits can accrue from administering punishment. Whether an observer is given this opportunity is randomly determined. Sometimes, none of the observers will be given the opportunity to punish the traders. Other times, only one observer will be given this opportunity. And, still, other times, both observers will be given this opportunity.

2.1.3. The first and second rounds

In the first and second rounds, real participants are assigned the role of traders. In both rounds, they are informed, at the end of each round, that another trader gave all of his or her endowment of 200 yen to them. Thus, in each round, participants make 400 yen plus the remaining portion of their own endowment of 200 yen that they did not give to another trader. In the first round, participants decide how much of their endowment of 200 yen to give to another participant without knowing what other traders will do. In the second round, they have an experience of receiving 400 yen from another trader in the first round before making the giving decision. It is well documented that game players do not cooperate when they expect that other players will defect (and thus, they will not receive benefits from other
players’ cooperative behavior; see Ostrom, 2003, Yamagishi, 1995, for reviews of the effects of expectations on cooperation). Thus, participants’ defection in the second round is less likely to reflect a fear that their cooperation will be exploited by others than it is in the first round. Instead, participants’ defection in the second round is more likely to reflect their willingness to exploit others to further their own interest. Therefore, we use the participants’ behavior in the second round to classify them as cooperators and noncooperators.

2.1.4. The third round

In the third round, participants are assigned the role of an observer. After the six traders make their decisions, the observers are informed that they are being given the opportunity to punish traders. They are first informed of how much each of the six traders gave. The three traders of Group A gave 50, 180, and 200 yen. The three traders of Group B gave 120, 160, and 200 yen. Which of the two groups is the participant’s own group is manipulated as a between-subjects factor. In the in-group cheater condition, Group A, containing the cheater, or the least cooperative member, who gave only 50 yen, is the in-group, while Group B is the out-group. In the out-group cheater condition, Group B is the in-group, while Group A, containing the cheater, is the out-group. Participants were allowed, as observers, to spend up to 200 yen to punish each of the six traders. The participants pay from their own earnings (including the initial show-up fee) to punish traders. Thus, punishment is costly to the punisher. Three times as much money as the participant pays to punish a trader is subtracted from the earnings of the punished trader. Participants are told that this is the third and the last round. This knowledge makes the punishment costly and deemed altruistic behavior, as discussed earlier. There is absolutely no possibility that the punishment affects participants’ long-term benefits either through altering other participants’ behavior in the future or acquiring a reputation that may be beneficial in the future (cf. Barclay, 2004).

2.2. Design

The experiment is a $2 \times 6$ design. The first factor involves the group membership of the cheater or the least cooperative trader: the in-group and the out-group cheater conditions. The second factor involves the six traders who are targets of punishment. The membership of the cheater (in-group and out-group cheaters) is a between-subjects factor and the second factor involving the six traders is a within-subjects factor. Participants decide how much of their own money they want to use to punish each of the six traders. In the in-group cheater condition, the three in-group traders gave 50, 180, and 200 yen, respectively, while the three out-group traders gave 120, 160, and 200 yen, respectively. In the out-group cheater condition, the amounts that in-group and out-group members give are reversed.

2.3. Participants

Ninety (52 males, 38 females) undergraduate students from a subject pool at Hokkaido University in Japan participated in this experiment. The potential participants in the subject pool were undergraduate students who registered in response to solicitation for participation.
in psychological experiments. The solicitation occurred in various classes. Monetary rewards were emphasized while recruiting potential participants. No class credits were involved. Three male participants were eliminated from the analyses because they were found to be suspicious about the experimental procedure or failed to understand the procedure. (Two participants answered positively to the question in the postexperimental questionnaire: “I suspected seriously that my earnings from the experiment had been previously determined by the experimenter and not by the interaction between myself and other participants in the experiment.” One participant stated that he made a mistake in deciding whom to punish.)

3. Findings

3.1. Cooperation in the gift-giving game

Consistent with findings in previous studies, considerable cooperation toward in-group members was observed in the first two rounds of the gift-giving game. (Note that participants were given opportunities to give only to another in-group member. They did not have a chance to give to out-group members.) Only two of the 87 participants gave nothing. The average amount of endowment that the participant gave was 125.84 yen (S.D. = 62.91) or 62.9% of the endowment in the first round and 147.02 yen (S.D. = 65.52) or 73.5% of the endowment in the second round. Cooperation toward in-group members was significantly higher in the second than in the first round \( t(86) = 3.18, p < .01 \). There was no significant difference between male and female participants in terms of the amount of endowment that they gave in either round. The fact that participants gave more in the second round suggests that participants reciprocated, in an indirect manner, the cooperation of an in-group member who gave them the entire amount of the endowment (200 yen) in the first round. In other words, the expectation of cooperation seems to have motivated cooperation. Actually, more than half of the participants (45 of the 87 participants) fully reciprocated; they gave an in-group member their entire endowment (200 yen) in the second round. The remaining participants gave, on average, only half of their endowment in the second round. The average amount of the endowment that these participants gave was 90.26 yen (S.D. = 51.21). In the analysis presented below, we refer to the participants who gave all of their endowment in the second round as cooperators, and the rest of the participants as noncooperators. In addition to the difference in their giving in the second round, cooperators and noncooperators differed greatly in their beliefs about other in-group members’ cooperativeness. Specifically, the average response (on a scale ranging from 1, \textit{did not think so at all}, to 9, \textit{strongly thought so}) to the question, “To what extent did you believe that \textit{IF} you gave your endowment to a student in your own \textit{gakubu}, he or she would, in return, give you more of his or her endowment?” was significantly higher among cooperators (6.80, S.D. = 2.47) than among noncooperators [4.02, S.D. = 2.47; \( t(85) = 5.23, p < .0001 \)]. Furthermore, as will be presented later, these two types of participants administered punishment to in-group and out-group members differently.
3.2. Third-party punishment

Punishment is not only costly for the punished individual but also for the third-party punisher. If the participants cared only about their own monetary status, then punishment would not occur. In sharp contrast to this prediction, which is based on a self-regarding model of human beings, a sizeable percentage of participants spent their money to punish. Fifty-three participants (about 61%) paid some cost to punish when they were in the role of observers. The average amount of money that participants spent for punishing was 64.91 yen (S.D. = 80.81). Most of the punishment was directed to the cheater (the least cooperative trader), but some punished more cooperative traders as well (see Fig. 1). When the target of punishment was the least cooperative trader (who gave only 50 yen of their endowment), participants spent an average of 34.31 yen to punish him or her (S.D. = 40.80). As a result, the least cooperative trader was fined by 102.93 yen, which is 69% of the endowment that the least cooperative traders kept to themselves.

Fig. 1 clearly indicates that most of the punishment was directed to the least and, to a lesser degree, the second least cooperative trader, both by cooperators and by non-cooperators. The figure represents the average proportion (of the sum of endowment they used to punish the six traders) that individual participants used to punish each trader. We used this proportional measure of punishment rather than the raw amounts that participants spent for punishing each trader because the total amount that they spent for punishment varied among participants. The use of raw amounts would give participants who used a lot of money for punishment larger weights in calculating the mean punishments for the six targets. In other words, the proportional measure of punishment produces a nonweighted rather than a weighted mean, weighted by the total amount that participants spent on punishment. (The proportion measure of punishment was set at zero for all targets for the

![Fig. 1. Average percentage of the money spent by cooperators and noncooperators for punishing each of the six traders.](image-url)
participants who spent no money at all for punishment. An alternative analysis excluding those participants produced basically the same results.) The pattern shown in Fig. 1 is consistent with the finding of Fehr and Fischbacher (2004).

3.3. In-group versus out-group punishment

We predicted that third-party punishment would be directed toward in-group members more often (and more severely) than toward out-group members because we regard punishment as second-order cooperation. Furthermore, we predicted that the in-group punishment would occur only among in-group cooperators and not among in-group noncooperators. As shown in the left panel of Fig. 1, this hypothesis was supported among cooperators who gave all of their endowment in the second round. On average, these participants spent 36.4% of their punishment expenditure to punish in-group members and 23.6% to punish out-group members. (These two numbers do not sum to 100% because some participants did not punish either in-group or out-group members.) The main effect of the target group in the Cheater Group × Target Group ANOVA was marginally significant \[ F(1,43) = 3.11, \ p < .09 \]. Furthermore, the same figure indicates that the difference in punishment between the in-group and out-group is more pronounced when the target is the least cooperative trader. The least cooperative in-group trader received 50.1% of punishment, whereas the least cooperative out-group trader received only 25.0%. This difference is statistically significant \[ t(43) = 2.13, \ p < .05 \]. These results support the prediction that cooperators punish in-group cheaters more than they punish out-group cheaters.

We do not make any prediction about the group difference in punishment among noncooperators. Interestingly, the result of our experiment among noncooperators, shown in the right panel of Fig. 1, indicates that out-group punishment is a mirror image of the in-group punishment that occurred among cooperators. Noncooperators, on average, spent 39.9% of punishment money to punish out-group members, compared with 22.0% on in-group members, and the difference is statistically significant \[ F(1,40) = 6.22, \ p < .05 \]. Furthermore, nonreciprocators’ tendency to punish out-group members was more pronounced when the target was the least cooperative trader. Least cooperative traders received 45.1% of punishment when they were an out-group member, whereas they received 22.0% when they were an in-group member. The difference was statistically significant \[ t(40) = -2.16, \ p < .05 \]. These findings, on the one hand, support our prediction of in-group punishment among cooperators. On the other hand, they illustrate an interesting pattern of out-group punishment among noncooperators. In the next section, we discuss how to interpret the serendipitous finding of out-group punishment by noncooperators.

3.4. Motivations behind third-party punishment

The proximate cause of punishment of cheaters is the negative emotion of anger or moral outrage as shown by Fehr and Gächter (2002). Evolutionary psychologists emphasize the role of punitive sentiments toward others in solving the collective action problem (Price et al., 2002). If the proximate cause of punishing behavior observed in our experiment is the
emotion of anger toward cheaters in the giving game, then we should find a correlation between the amounts of money participants spent for punishing cheaters and the level of anger that they felt toward cheaters. Furthermore, we believe that participants who punish cheaters are those who espouse the state of mutual cooperation as a goal and thus might feel guilty for not punishing cheaters and thus failing to promote mutual cooperation. To see if participants’ punishment behavior is related to their emotional responses to the cheaters, we measured their emotional responses to the cheaters with the following questions in the postexperimental questionnaire:

- How strongly did you feel anger toward the members who didn’t give much of their endowment to another member of (your or another) gakubu? (anger);
- How strongly did you feel a sense of unfairness toward the members who didn’t give much of their endowment to another member of (your or another) gakubu? (unfairness);
- How strongly would you have felt sorry for those in (your or the other) gakubu who were not given money if you did not punish those who did not give? (guilt).

Participants responded to these six questions using a nine-point response scale (1 = not at all, 9 = very much).

Table 1 reports the correlation coefficients between the participants’ responses to these items and the proportion of money that they spent punishing in-group and out-group cheaters. Feelings of anger, unfairness, and guilt are all positively correlated with the level of punishment of the in-group cheater (i.e., the least cooperative member), consistent with Fehr and Gächter (2002). These correlations were absent with regard to out-group cheaters. These correlations suggest that the same behavior (i.e., reducing another individual’s income) may be based on different motives when it is directed to the in-group than when it is directed toward the out-group. Participants reduce in-group members’ income as an emotional response toward cheaters. In other words, they punish cheaters who incite moral outrage. When cheaters are in another group, the same behavior does not seem to be based on negative emotions toward cheaters.

4. Conclusion

A number of studies in social psychology and experimental economics have shown that cooperative behavior is more likely to be directed toward in-group members than it is to be
directed toward out-group members. We extended the logic of the group-based nature of cooperation to include punishment as second-order cooperation. We conducted an experiment to empirically demonstrate that punishment is group based in the same way that cooperation is. The experimental findings support our hypothesis that those who cooperate within a group will punish in-group members more than they will punish out-group members. Postexperimental questionnaire data add more support to our hypothesis. Specifically, negative emotions toward cheaters are related to the punishment level of in-group members but are not related to the punishment level of out-group members. Our findings that punishment is mostly directed toward in-group rather than out-group members suggest that the group-based nature of cooperation is extended to punishment as second-order cooperation.

While we did not have specific predictions about the punishment pattern of non-cooperators, our experiment produced an interesting finding. Noncooperators punished out-group cheaters more strongly than they did in-group cheaters. Postexperimental questionnaire data indicate that the punishment of out-group cheaters is not based on the participants’ moral outrage toward them. That is, the questionnaire data show that the punishment of the out-group is not related to emotions such as anger or unfairness toward free riders. This result is consistent with a finding by Masclet, Noussair, Tucker, and Villeval (2003). They report that nonmonetary disapproval of others’ decision in a public-goods game is directed toward only noncooperators, while monetary punishment is directed toward both noncooperators and cooperators. Their findings suggest that monetary punishment may involve two motivational bases: moral outrage against norm violators and competitive social motivation. Our questionnaire data indicate that the major motivational basis for in-group punishment is moral outrage, whereas moral outrage is not related to out-group punishment. Masclet et al. (2003) suggest that competitive social motivation (Falk, Fehr, & Fischbacher, 2001; Messick, & McClintock, 1968; Saijo & Nakamura, 1995) may explain the punishment of cooperators. That is, participants in their experiment reduced both cooperators’ and noncooperators’ earnings to enhance the relative standing of the self against others. We speculate that moral outrage is the dominant motivational bases of in-group punishment, whereas competitive social motivation is the dominant motivational basis of reducing out-group members’ earnings. A support of this interpretation comes from another experiment by Shinada, Yamagishi, & Yamamoto (unpublished). Punishment of the out-group is useless as a means to enhance one’s relative standing against the out-group when the cost of the punishment is the same as the benefit of the punishment. In our experiment, the cost/benefit factor is 3. That is, three times the cost of punishment is subtracted from the earnings of the punished. This ratio implies that punishment can be a means to enhance one’s standing against the out-group. However, if the cost/benefit factor is reduced to 1, the punisher loses as much as the punished does, and thus, the punisher cannot improve his or her relative standing against the out-group by punishing an out-group member. Shinada et al. (unpublished) replicated the current experiment, in which the cost/benefit ratio is reduced to 1. As expected, out-group punishment was greatly reduced, whereas in-group punishment was only marginally affected by the reduction of the cost/benefit factor to 1.

The finding that in-group and out-group punishments involve different motivational bases has important implications to the study of altruistic and third-party punishment. The finding is
important, in part, because research on punishment behavior in this tradition has implicitly assumed that punishment is a means to redress inequity and unfairness. This may be true for in-group punishment, but not true for out-group punishment. Future studies of altruistic punishment are encouraged to introduce an experimental design that can clearly separate punishment from spiteful behavior.

Another important implication of the findings of this experiment is the ramifications of punishment and spiteful behavior in ongoing inter- and intragroup interactions. The “punished” cheaters may respond differently to in-group and out-group punishment, while the two take exactly the same form of income reduction. The harmful acts of reducing the cheater’s welfare may be interpreted by the cheater as an act of punishment. The best way to avoid being punished is to acquiesce to the demands of the punishers and cooperate. Thus, in-group punishments promote in-group cooperation (Fehr & Gächter, 2002; Ostrom, Walker, & Gardner, 1992; Yamagishi, 1986b, 1988). The same harmful acts that come from out-group members, on the other hand, are not likely to be interpreted as a means to enhance cooperation because cooperation by the punished do not help out-group members. Thus, the harmful acts are likely to be interpreted as aggression rather than punishment. The difference in the interpretations of the same harmful acts might thus provide very different long-term consequences in inter- and intragroup interactions. Punishment may promote cooperation within groups, while it may promote conflicts between groups. Mixing punishment with spiteful behavior may be the best way to turn goodwill into antagonism.

Acknowledgments

The research reported in this paper has been supported by a grant from the Japan Society for the Promotion of Science. We thank colleagues at Hokkaido University who helped us recruit potential participants from their classes.

References

Dawes, R. M., van de Kragt, A. J. C., & Orbell, J. M. (1988). Not me or thee but we: the importance of group


