Within-individual associations among third-party intervention strategies: Third-party helpers, but not punishers, reward generosity

Ohtsubo, Yohsuke
Sasaki, Shunta
Nakanishi, Daisuke
Igawa, Junichi

(Citation)
Evolutionary Behavioral Sciences,12(2):113-125

(Issue Date)
2018-04

(Resource Type)
journal article

(Version)
Accepted Manuscript

Within-Individual Associations among Third-Party Intervention Strategies: Third-Party Helpers, but Not Punishers, Reward Generosity

Yohsuke Ohtsubo Shunta Sasaki

(Kobe University)

Daisuke Nakanishi

(Hiroshima Shudo University)

Junichi Igawa

(Oita University)

Author Note

Yohsuke Ohtsubo, Department of Psychology, Graduate School of Humanities, Kobe University, Japan; Shunta Sasaki, Department of Psychology, Graduate School of Humanities, Kobe University, Japan; Daisuke Nakanishi, Faculty of Humanities and Human Sciences, Hiroshima Shudo University, Japan; Junichi Igawa, Faculty of Economics, Oita University, Japan.

We are grateful to Thomas McCauley for his valuable comments. This research was supported by the Japan Society for the Promotion of Science (No. 15KT0131).

Correspondence concerning this article should be addressed to Yohsuke Ohtsubo, Graduate School of Humanities, Department of Psychology, Kobe University, Kobe, 657-8501, Japan. E-mail: yohtsubo@lit.kobe-u.ac.jp

Abstract

Third parties intervene in others' behaviors in various ways, such as punishing a harm-doer and/or helping a victim. Moreover, third parties may reward generous altruists. As such, various types of third-party intervention strategies are conceivable. Nevertheless, researchers have disproportionately focused on third-party punishment. In the present study, 87 undergraduate students were exposed to unfair, fair, and generous allocators and their recipients, and were allowed to punish, help, and/or reward the players. The results indicated that participants were more likely to punish unfair allocators than fair allocators; more likely to help recipients of unfair allocators than those of fair allocators; and more likely to reward generous allocators than fair allocators. Examinations of intra-individual associations of these strategies revealed that two prosocial strategies (third-party help and third-party reward) were strongly tied to each other (i.e., participants who helped victims of unfair allocators were more likely to reward generous allocators). However, third-party punishment was not significantly associated with the other two strategies. The emotional correlates of the three intervention strategies were also investigated. Third-party punishment was correlated with moral outrage and reduced empathic concern for unfair allocators. Third-party help was correlated with empathic concern for the victim. Third-party reward was correlated with empathic concern for generous allocators.

Keywords: third-party punishment, third-party help, third-party reward, moral outrage, empathic concern

Within-Individual Associations among Third-Party Intervention Strategies: Third-Party Helpers, but Not Punishers, Reward Generosity

Punishment is defined as a behavior that reduces non-cooperators' fitness to promote their cooperation in future interactions. By imposing punishment, punishers themselves typically incur some immediate costs (i.e., reduce their own fitness). But benefits of punishment accrues from the punished party's more cooperative behavior in the future (Clutton-Brock & Parker, 1995; Raihani, Thornton, & Bshary, 2012). Although punishment that meets this definition is relatively rare in non-human animals (Raihani et al., 2012), humans critically rely on punishment to maintain large-scale cooperation (Fehr & Fischbacher, 2004a; Gintis, 2000). Collective management of common natural resources is an example of large-scale cooperation (Hardin, 1968). Suppose that in order to keep a forest sustainable, a community sets a rule regarding the maximum amount of wood that each community member can harvest each year. However, each member has an incentive to violate the rule because he/she can be better off by overharvesting wood to increase his/her personal income. Field research has shown that the presence of punishment against non-cooperators (i.e., those who have violated the rule) predicts successful common resource management (Ostrom, 2000; Rustagi, Engel, & Kosfeld, 2010). Experiments simulating this type of cooperation problem in laboratories have also shown that the cooperation rate increases by allowing group members to punish non-cooperative members (Fehr & Gächter, 2002; Yamagishi, 1986).

People's punitive tendency is also epitomized by so-called third-party punishment, where people punish non-cooperators even when they themselves were not affected by the non-cooperators' behavior. In the standard third-party punishment game (Fehr & Fischbacher, 2004b), a participant observes the dictator game played by two players (the allocator and the

recipient). The allocator is given a certain amount of endowment and is asked to divide it between him/herself and the recipient. The allocator can divide the resource as he/she likes, and thus behave like a dictator. Observing the allocator's decision, the participant (the third party) is asked to decide whether to spend some of his/her own endowment to reduce the allocator's payoff. If the third-party spends *x* game points, *cx* points will be subtracted from the allocator's payoff (where *c* is typically 2 or 3). In Fehr and Fischbacher's (2004b) seminal experiment, the allocator's unfair behaviors (e.g., giving only 10% of the endowment to the recipient) elicited costly punishment from the third-party participants. Third-party punishment has been observed in many subsequent experiments that varied in various ways, such as the experimental method to measure the third-party responses, types of the norms that were violated, and frequency of non-cooperators (e.g., Bone, Silva, & Raihani, 2014; Jordan, McAuliffe, & Rand, 2016; Konishi & Ohtsubo, 2015; Nelissen & Zeelenberg, 2009; Ohtsubo, Masuda, Watanabe, & Masuchi, 2010). Moreover, when tested using population-adjusted versions of the third-party punishment game, young children (e.g., McAuliffe, Jordan, & Warneken, 2015; Riedl, Jensen, Call, & Tomasello, 2015) as well as people in small-scale societies (e.g., Henrich et al., 2006, 2010; Marlowe et al., 2008) have inflicted third-party punishment on non-cooperators. Therefore, humans seem inherently punitive.

However, it is noteworthy that third-party punishment experiments have typically only allowed participants to punish non-cooperators. Thus the apparent prevalence of a punitive tendency might be due to procedural constraints (Pedersen, Kurzban, & McCullough, 2013). Outside the laboratory, punishment is only one intervention strategy that third parties can employ. Third parties might help victims, instead of punishing violators. If the goal of punishment is to promote cooperation, it is also possible for third parties to reward cooperators, instead of

punishing non-cooperators. Although research on third-party intervention strategies has disproportionately focused on punishment, interest in other intervention strategies (i.e., third-party help and third-party reward) seems to be emerging (Almenberg, Dreber, Apicella, & Rand, 2011; Charness, Reyes, & Jiménez, 2008; Gummerum, Van Dillen, Van Dijk, & López-Pérez, 2016; Hu, Strang, & Weber, 2015; Leliveld, van Dijk, & van Beest, 2012; Lotz, Baumert, Schlösser, Gresser, & Fetchenhauer, 2011; Lotz, Okimoto, Schlösser, & Fetchenhauer, 2011).

However, in the case of third-party help and reward, the question arise as to why third parties incur costs to increase unrelated others' well-being? A recently proposed adaptive function of third-party help is the signal of trustworthiness (Jordan, Hoffman, Bloom, & Rand, 2016). In fact, the costly signaling hypothesis of third-party intervention was originally developed in the context of third-party punishment (e.g., Barclay, 2006; Kurzban, DeScioli, & O'Brien, 2007). According to this hypothesis, third-party punishment is a costly signal of punishers' cooperativeness. However, empirical support for this hypothesis has been, at best, mixed (see Raihani & Bshary, 2015a, for a review). Punishers in fact tend to earn a cooperative reputation (e.g., Barclay, 2006), which, however, does not necessarily fare better than the reputations of other types of players, such as third-party helpers and rewarders (Horita, 2010; Kiyonari & Barclay, 2008; Ozono & Watabe, 2012; Raihahi & Bshary, 2015b).

Raihani and Bshary (2015a) pointed out that punishment can signal not only the punishers' cooperativeness but their competitiveness as well. Accordingly, when other more prosocial options, such as helping and rewarding, are available, punishment becomes a less credible signal of third parties' cooperativeness. Jordan, Hoffman, et al. (2016) tested the validity of this modified version of the costly signaling hypothesis of third-party punishment by

conducting an experiment that combined the third-party punishment game with the trust game. In the experiment, there were four roles: the three roles in the standard third-party punishment game (i.e., allocator, recipient, third party) and the observer of the third-party punishment game. After learning whether or not the third party punished the unfair allocator in the third-party punishment game, the observer played the trust game with the third party. In particular, the observer decided what proportion of his/her endowment he/she would transfer to the third party (the second mover in the trust game). The transferred money was tripled and given to the second-mover. The second-mover then decided how much money to send back to the observer. Without the help option, third-party punishment was in fact a signal of the third-party players' cooperativeness: The observers trusted punishers more than non-punishers, and the punishers actually behaved in a more trustworthy manner than the non-punishers. However, once the third-party help option was introduced, the observers preferred helpers to punishers. Moreover, when both punishment and help options were available, punishers no longer behaved in a trustworthy manner (i.e., they did not return a fair share of their partner's money), while helpers behaved in a more trustworthy manner. These results indicate that third-party punishment is a less credible signal of cooperativeness than third-party help (a more prosocial intervention strategy), especially when players can choose to either punish a perpetrator or help a victim.

The modified version of the signaling hypothesis of third-party intervention presumes that prosocial intervention strategies, such as third-party help and reward, are more credible signals of cooperativeness than third-party punishment (Jordan, Hoffman, et al., 2016; Raihani & Bshary, 2015a). It is interesting to note that non-negligible portions of participants in previous studies produced a more credible signal (i.e., third-party help or reward) in combination with the less credible signal (i.e., third-party punishment). For example, in Lotz, Okimoto, et al.'s (2011)

experiment, 42.6% of participants used both punishment and help options (32.0% used the help option only and 6.7% used the punishment option only). Similarly, when participants were given the punishment and reward options in Almenberg et al.'s (2011) experiment, 36% of participants not only punished greedy allocators but also rewarded generous allocators (20% used the reward option only and 9% used the punishment option only). Given these findings, the question arises as to why a substantial portion of participants did not concentrate their resource on the more credible signal (i.e., help or reward), but instead split their resource between the two options. Participants might have attempted to enhance the credibility of the signal by assuming that observers would perceive someone who performs both (i.e., third-party punishment and help/reward) as being more cooperative. This reasoning led to the prediction that when given all possible options (i.e., punish, help, and reward), participants would combine the most credible signals (i.e., help and reward), instead of combining one credible signal (i.e., help or reward) and one less credible signal (i.e., punish). Therefore, it was hypothesized that those who help a victim are also more likely to reward a generous allocator, while they are not necessarily more likely to punish a greedy allocator. This hypothesis is divided into the following two components.

Hypothesis 1a: *The intra-individual association between "help" and "reward" is stronger than the association between "punish" and "help."*

Hypothesis 1b: *The intra-individual association between "help" and "reward" is stronger than the association between "punish" and "reward."*

In no previous study were participants given all possible behavioral options at once. Therefore, in the present study, all options were made available to third-party participants to test the above hypotheses.

In addition, we explored the emotional correlates of these intervention strategies.

Research has shown that moral outrage (or indignation) causes third-party punishment (Fehr & Fischbacher, 2004a; Gummerum et al., 2016; Jordan, McAuliffe, et al., 2016). Lotz and colleagues also examined the effect of moral outrage and found that it predicted not only third-party punishment but also third-party help (Lotz, Baumert, et al., 2011; Lotz, Okimoto, et al., 2011). Leliveld et al. (2012) examined the effects of empathic concern (Davis, 1983), and revealed that participants high in trait empathic concern tended to help victims, while participants low in trait empathic concern tended to punish unfair players (see also Hu et al., 2015; Konishi, Oe, Shimizu, Tanaka, & Ohtsubo, in press). The effect of empathic concern on third-party help is consistent with evolutionary theories positing that compassion, which is conceptually similar to empathic concern, facilitates prosocial responses to someone's suffering (Goetz, Keltner, & Simon-Thomas, 2010). This effect is also consistent with the well-established finding that empathy promotes altruistic behavior in general (Batson, 1991; Batson, Duncan, Ackerman, Buckley, & Birch, 1981). To our knowledge, there has been no research directly examining the emotional correlates of third-party reward. However, it is plausible that empathic concern for generous allocators promotes third-party reward, as this is also a form of altruistic behavior. Based on the aforementioned previous studies, we formulated the following three hypotheses:

Hypothesis 2a: *Moral outrage and reduced empathic concern for the unfair allocator are correlated with third-party punishment.*

Hypothesis 2b: *Moral outrage and empathic concern for the victim are correlated with third-party help.*

Hypothesis 2c: *Empathic concern for the generous allocator is correlated with third-party reward.*

In sum, the primary purpose of the present study was to investigate the intra-individual associations among third-party intervention strategies. In particular, we hypothesized that third-party help and third-party reward would be more strongly tied together as compared to the association between third-party punishment and each of the two strategies. The secondary purpose was to examine the emotional correlates of each of the three intervention strategies. We particularly focused on anger at the non-cooperators (i.e., moral outrage) as well as empathic concerns for unfair allocators, their victims, and generous allocators.

Method

Participants and the Third-Party Intervention Game

This study involved two separate groups of undergraduate students. The first group of 11 undergraduate students (7 females and 4 males) played the dictator game with each other. The second group of 87 undergraduate students (49 females, 38 males, M_{age} = 19.87 years, SD = 0.85) played the third-party role.

The rule of the third-party intervention game was as follows: Two players first play the dictator game (throughout the instructions, neutral language, such as the allocation game, allocator, and recipient, were used). The allocator receives 100 Japanese yen (100 JPY ≈ US$ 0.80) as his/her endowment and decides how to allocate it between him/her and the recipient (with 5 JPY as the minimal unit of the allocation). The third-party player receives 100 JPY as his/her endowment. The third-party player is allowed to make any interventions he/she would like using his/her endowment. The intervention options comprise “to increase, decrease or do nothing about the allocator’s payoff” and “to increase, decrease or do nothing about the recipient’s payoff.” If the third-party player chooses to either increase or decrease the other players’ payoffs, he/she has to spend some of his/her endowment. The effect of the third-party

player's spending on the target's payoff is twice as much as its effect on his/her own payoff. For example, if the third-party player decided to spend *x* JPY to decrease (or increase) the allocator's payoff and spend *y* JPY to increase (or decrease) the recipient's payoff, 2*x* JPY is subtracted from (or added to) the allocator's payoff and 2y JPY is added to (or subtracted from) the recipient's payoff. In this example, the third-party player's payoff would be $(100 - x - y)$ JPY. The third-party player is allowed to increase/decrease the payoffs of both players, one of the two players, or none of the two players. Although the third-party player is allowed to spend any amount he/she wants (in increments of 5 JPY), it is not permitted to increase either player's payoff beyond 100 JPY or reduce either player's payoff below zero.

Notice that there was an asymmetry in the payoff between the third-party and the other two players. The allocator divided 100 JPY (i.e., the average payoff of the allocator and recipient was 50 JPY), while the third-party received 100 JPY for him/herself. We decided to give twice as much (i.e., 100 JPY) to the third-party participants so that their behaviors would not be restricted by any budgetary concerns (e.g., if the third-party participants were given only 50 JPY, for example, they might avoid being worse off than an unfair allocator by helping his/her recipient). In addition, this precluded envy from causing apparent punishment-like behavior (cf. Pedersen et al., 2013).

The Task of the First Set of Participants

The first set of 11 participants played the allocation game with each other. Since the 11 participants knew each other before the experiment, they were told that their decisions would be kept strictly confidential to their partners. In addition to the above rules of the third-party intervention game, they were told that their photographs and their decisions would be shown to the second set of third-party participants (students of a different university in a distant city). Each

of the 11 participants played the allocation game with every one of the other 10 participants as the allocator. Therefore, each participant was involved in 20 allocation games (10 games as the allocator and 10 games as the recipient). They were told that their reward for this study would be the sum of their payoffs in all 20 third-party intervention games.

The first set of participants were given 100 JPY for each round of the allocation game in which they played the allocator role. They decided how to allocate the endowment between themselves and their partner. Although the first set of participants were paid according to their actual payoffs, they took part in this study as a part of their psychology course. Accordingly, we explicitly asked them to include various allocations, such as unfair, fair, and generous allocations, in their 10 decisions, so that they would learn how others would react to different types of allocations. Therefore, their role in the experiment was somewhat similar to confederates. After making the 10 allocation decisions, they were individually photographed so that their pictures would be associated with their decisions in the later third-party intervention experiment. We decided to present these photographs to the third-party participants to enhance the reality of the experiment.

The Task of the Second Set of Participants

The second set of participants played the role of the third-party. Since the primary interest of this study was third-party intervention, in the subsequent sections, we refer to this second group as “participants.” Each of the participants was presented with five allocations made by the first set of participants, and decided whether to intervene in five allocation game results (the five allocation games always involved 10 different players). Although we explained the rule of the third-party intervention game to participants, we did not inform them that the first set of participants had known each other before the experiment, had been involved in multiple rounds

of the allocation game, and had been asked to include various allocations in their allocation decisions. Participants were told that they would earn the sum of the money that they did not spend on intervention in the five allocation games.

The five allocations consisted of fair allocations (giving the partner 45, 50, or 55 JPY out of 100 JPY), unfair allocations (giving less than 45 JPY to the partner), and generous allocations (giving more than 55 JPY to the partner).[1] The five allocations always included at least one fair, one unfair, and one generous allocation. Since there were five allocations, two of the three types of allocations were duplicated. However, unless otherwise noted, each participant's first response to each type of allocation was analyzed. The two other allocations were treated as fillers. We included the two filler allocations in order to minimize participants' suspicion—if they were exposed to only three different types of allocations (i.e., fair, unfair, generous), they might unnecessarily suspect deception.

Procedure

The main part of the study was conducted as part of larger data collection sessions. Each session involved four to 10 participants. Participants signed a consent form covering all studies included in the session. They first filled out a packet of individual differences questionnaires. After all participants completed the first packet of questionnaires, the experimenter explained the nature of the present study by projecting instruction slides on a large screen. The instructions read as follows: Other participants at a different university engaged in a series of allocation games. Participants will see the results of five different pairs in the allocation game. Their task is to make a series of five decisions regarding whether to spend some of their endowment (100 JPY for each decision) to modify each of the five allocations. In particular, for each allocation game, the pictures of the allocator and the recipient are presented on the left- and right-hand sides of the

screen, respectively, along with the allocator's decision. Each picture is accompanied by a response box in which participants choose to either increase, decrease, or do nothing about the corresponding player's payoff. If they choose to increase or decrease it, they are asked to determine how much they will spend to do so.

After confirming that participants understood the above instructions, they moved to another experimental laboratory. In the laboratory, there were five separate cubicles, each equipped with a laptop computer. When fewer than six participants were involved in the session, all participants moved and engaged in the third-party intervention experiment simultaneously. When there were more than five participants, half of them first moved and engaged in this study, while the remaining participants stayed in the original room and filled out the second set of questionnaires. Once the first half of participants completed the experiment, the remaining participants then moved to the laboratory and engaged in the third-party intervention experiment.

In the laboratory, participants individually took part in this experiment. They made the intervention decisions for each of the five allocations consecutively. First, they were shown one allocation decision accompanied by the pictures of a pair of players, and they decided whether to change the allocator's payoff (and if so, how much) and whether to change the recipient's payoff (and if so, how much). To confirm participants' understanding, we asked them to enter the final payoffs of the three players (the allocator, the recipient, and themselves) in three corresponding boxes that appeared on the screen right after they made the two decisions. Once participants finalized their intervention decisions, they were presented with 12 emotional words and were asked to indicate how strongly they felt each of the 12 emotions for the allocator on a 6-point scale (0 = "do not feel at all" to 5 = "very strongly feel"). These emotion items comprised five anger items (i.e., angry, indignant, mad, outraged, perturbed), five empathic concern items (i.e.,

sympathetic, compassionate, warm, tender, softhearted), and two envy items (i.e., envious, jealous). The anger items were adapted from Batson et al.'s (2007) study on anger at moral violations. The empathic concern items were adapted from Batson et al.'s (1981) study on the empathy-altruism hypothesis. The envy items were written by the authors. On the next screen, they were asked to rate how strongly they felt the same 12 emotions for the recipient on a 6-point scale. Participants repeated the same procedure five times. After completing the game experiment, participants moved back to the original room and completed the remaining part of the experimental session.

Participants received their monetary reward for the allocation game at the end of the experimental session. The rewards for the other parts of the session were paid into their bank account after a few weeks. To determine the reward for the first group of participants, we randomly chose some responses from those in the second group. The eleven participants in the first group were paid according to their own decisions and the intervention decisions of those in the second group. This experiment was approved by the research ethics committee of the Graduate School of Humanities, Kobe University.

Results

Basic Pattern of Strategy Uses

Third-party punishment, help, and reward were operationally defined as follows. If participants spent their resource to reduce the allocator's payoff in the unfair condition, it was considered an instance of third-party punishment. If participants increased the recipient's payoff in the unfair condition, it was considered an instance of third-party help. If participants increased the allocator's payoff in the generous condition, it was considered an instance of third-party reward. There were two additional possible responses. If participants reduced the recipient's

payoff in the generous condition, it was considered an instance of spite. If participants reduced the allocator's payoff in the generous condition, it was considered an instance of antisocial punishment (Herrmann, Thöni, & Gächter, 2008).

To visually inspect whether these strategies were employed in response to the expected allocations (i.e., third-party punishment and help in the unfair condition, and third-party reward in the generous condition), we computed the relative frequencies of the use of each option as a function of allocation. For this visual inspection, we included each participant's responses to all five allocations, which yielded 435 responses (see Figure 1). Since no participants in the first group chose one moderately unfair allocation (75/25 to allocator/recipient) and two extremely generous allocations (0/100 and 5/95), these three allocations do not appear in Figure 1. There are three prominent patterns in Figure 1. First, third-party punishment and help increased as allocations became unfair (see the right-hand side of Figure 1). Second, third-party rewards were prevalent when the allocator behaved in a generous manner (see the left-hand side of Figure 1). Third, participants rarely employed spite (only five of 87 participants reduced the recipient's payoff in the generous condition) or antisocial punishment (no participants reduced the generous allocator's resource). The first two patterns are in line with the general notions of punishment, help, and reward.

Frequencies of Strategy Uses and Sex Differences

We then confirmed the validity of the above visual inspections by comparing the frequency of each of the three strategies in the fair (control) condition and the frequency in the relevant condition (i.e., the unfair condition for third-party punishment and help, and the generous condition for third-party reward). In response to the first unfair allocation that they saw, 25 of 87 participants (.29) decided to reduce the unfair allocator's payoff, whereas only three

participants (.03) did so in response to the first fair allocation. The difference was significant by McNemar's test, $\chi^2(1) = 20.05$, $p < .001$. For third-party help, 48 participants (.55) increased the recipient's payoff in the unfair condition, while 22 participants (.24) did so in the fair condition, $\chi^2(1) = 21.81$, $p < .001$. For third-party reward, 54 participants (.62) increased the allocator's payoff in the generous condition, while 23 participants (.26) did so in the fair condition, $\chi^2(1) = 25.71$, $p < .001$. These results indicate that third-party punishment, help, and reward were more frequently employed in the relevant condition than in the fair (control) condition. In contrast, for spite, only five participants (.06) decreased the recipient's payoff in the generous condition, and three participants (.03) did so in the fair condition, $\chi^2(1) = 0.13$, *ns*. This result implies that in this study, no meaningful employment pattern was observed for spite. As we noted, we observed no instances of antisocial punishment. Therefore, we did not include spite and antisocial punishment in the subsequent analyses.

We also compared whether some strategies were more likely to be employed than others. A series of McNemar's tests indicated that participants were more likely to help the victim of the unfair allocator (.55) than punish the unfair allocator (.29), $\chi^2(1) = 13.08$, $p < .001$; they were also more likely to reward the generous allocator (.62) than to punish the unfair allocator (.29), $\chi^2(1) = 19.12$, $p < .001$; however, they were not significantly more likely to reward the generous allocator (.62) than to help the victim of the unfair allocator (.55), $\chi^2(1) = 2.08$, $p = .149$.

The sex difference in the punishment rate (.32 and .27 for men and women, respectively) was not significant by Fisher's exact test, $p = .639$. However, the sex difference was significant for the help rate (.34 and .71 for men and women, respectively) and for the reward rate (.42 and .78 for men and women, respectively), $p < .001$ for both comparisons by Fisher's exact test.

We also measured the amount of the cost participants were willing to incur for each decision. However, third-party punishment and help were subject to the same budget constraint, as both were responses to the first unfair allocations (participants had to decide how to spend their 100 JPY endowment for punishment and help simultaneously), while third-party reward was independent of the constraint, as it was the response to the first generous allocation (participants decided how much of their 100 JPY they would spend almost solely on this intervention, given that few participants behaved in a spiteful manner to the recipients of generous allocations). Therefore, even if we observed more spending on reward than on the other two intervention strategies, it could have been due to either participants' preference or budget constraint. For this reason, we focused on the dichotomous variables indicating whether participants chose to use each intervention strategy, and we did not analyze the amount of cost.

Intra-individual Associations of Intervention Strategies

The intra-individual association between third-party punishment and help is shown in the top panel of Table 1; this association was marginally significant by Fisher's exact test, $p = .058$. The corresponding intra-individual correlation between punishment and help, designated as r(punish, help), was .21 (the correlations reported in this section are computationally equivalent to phi coefficients). The intra-individual association between third-party punishment and reward (shown in the middle panel of Table 1, r(punish, reward) = .18) was not significant by Fisher's exact test, $p = .142$. However, the intra-individual association between third-party help and third-party reward (shown in the bottom panel of Table 1, r(help, reward) = .72) was extremely high and significant by Fisher's exact test, $p < .001$.

Hypotheses 1a and 1b predicted that r(help, reward) would be greater than r(punish, help) and r(punish, reward). A series of Williams's tests for two correlated correlations

(performed using the r.test function in the psych package of R) indicated that *r*(help, reward) was significantly higher than *r*(punish, help), $t(84) = 5.02$, $p < .001$, and *r*(punish, reward), $t(84) = 5.47$, $p < .001$. Therefore, Hypotheses 1a and 1b were supported—participants were more likely to employ two prosocial intervention strategies together than to combine one prosocial strategy with the punitive strategy (i.e., third-party punishment).

Remember that there was a significant sex difference in the help and reward rates. In order to confirm that the above patterns would not be modified by the sex difference, we computed the same correlations for males and females separately: *r*(punish, help) was .23 and .28 for males and females, respectively (cf. it was .21 when the two sexes were combined); *r*(punish, reward) was .22 and .21 for males and females, respectively (cf. the combined $r = .18$); and *r*(help, reward) was .62 and .74 for males and females, respectively (cf. the combined $r = .72$). The significance pattern revealed by a series of Fisher's exact tests was the same except that the originally marginally significant association between punish and help became non-significant in the male sample.

Emotional Correlates of Each Strategy

We then computed point-biserial correlations among the three intervention strategies and emotional responses (see Table 2). Recall that participants indicated their levels of anger (5 items), empathic concern (5 items), and envy (2 items) for each of the two players (i.e., the allocator and the recipient). The anger and empathic concern items were associated with reasonable levels of Cronbach's alpha coefficients (see Table 2), whereas the alpha coefficients of the envy items were low (correlations between the two items ranged between .14 and .46). Therefore, envy scores in the present study are not reliable. However, to avoid problems of selective reports, we decided to include the correlations involving the envy score.

Confirming Hypothesis 2a, both moral outrage (i.e., anger at the unfair allocator) and reduced empathic concern for the unfair allocator were significantly correlated with third-party punishment. However, Hypothesis 2b was only partly supported. Although empathic concern for the victim was significantly correlated with third-party help, moral outrage was not significantly correlated with third-party help. The significant correlations relevant to Hypotheses 2a and 2b remained significant after controlling for the effect of sex (see the corresponding partial correlations in brackets in Table 2).

In the generous condition, we computed the point-biserial correlations between third-party reward and emotional responses. Confirming Hypothesis 2c, empathic concern for the generous allocator was significantly correlated with third-party reward. As shown in brackets, the partial correlation between empathic concern for the allocator and third-party reward remained significant even after controlling for the effect of sex. Unexpectedly, the two envy scores (associated with the generous allocator and his/her recipient) were significantly correlated with third-party reward. However, these correlations became non-significant after controlling for the effect of sex. Moreover, as we already noted, the reliability of the envy scores was low. Therefore, we refrain from interpreting significant correlations between third-party reward and the two envy scores.

Discussion

This study examined participants' responses to both unfair and generous allocations. For these two types of allocations, there are three meaningful third-party intervention strategies: punishing unfair allocators, helping the victims of unfair allocation, and rewarding generous allocators. The results indicate that participants employed these three intervention strategies more frequently in the relevant situation than in the irrelevant, control condition (i.e., in response

to the fair allocation). Moreover, those who helped the victim in the unfair condition were more likely to reward the allocator in the generous condition. This intra-individual association was stronger than the association between punishment and each of the two other strategies. This pattern is consistent with the costly signaling hypothesis of third-party intervention: participants signal their cooperativeness by intervening in a resource allocation in which they are not directly involved. Raihani and Bshary (2015a) pointed out that prosocial intervention strategies are more credible signals of cooperativeness because punishment, which is a less prosocial strategy, can signal both cooperativeness and competitiveness. Jordan, Hoffman, et al. (2016) found that participants actually endorsed third-party help more than third-party punishment as a signal of cooperativeness. This study extended this finding by showing that third-party helpers were more likely to use another prosocial strategy—namely, third-party reward. Apparently, prosocial people attempt to enhance signal credibility by combining the two prosocial signals.

We also investigated the emotional correlates of each of the three intervention strategies. As predicted from previous studies, third-party punishment was correlated with moral outrage and reduced empathic concern for the unfair allocator. Empathic concern for the victim was correlated with third-party help. Although these patterns corroborate previously reported findings, we failed to replicate one interesting finding. Gummerum et al. (2016) found a significant correlation between third-party help and empathic anger at the unfair allocator. Admittedly, we did not measure empathic anger separately from “overall” anger at the unfair allocator, which might have comprised empathic anger and moral outrage (see Batson, 2011 for the differences between these two emotions). In future studies, the effects of moral outrage, which supposedly promotes punishment, and empathic anger, which may promote help for the victim, should be separated and more carefully examined.

In addition, consistent with the empathy-altruism hypothesis (Batson, 1991), empathic concern for the generous allocator promoted third-party reward. Although we used the same empathic concern items across targets and conditions, empathic concern for generous allocators is plausibly associated with moral elevation and the feeling of warmth, while empathic concern for victims is conceivably associated with compassion, sympathy and pity (Goetz et al., 2010). These are likely distinct emotional responses, and future studies need more nuanced measures of the different aspects of empathic responses.

The costly signaling hypothesis of third-party intervention predicted that participants would use prosocial strategies, and abandon the ambiguous signal (i.e., third-party punishment). Nevertheless, a non-negligible portion of participants (.29) still inflicted punishment. In addition, we observed that third-party punishment and help had different sets of emotional correlates. Therefore, it seems premature to conclude that every punitive intervention observed under the restricted intervention options can be subsumed under the cooperative signal. This result suggests that at least some percentage of people punish non-cooperators simply for the sake of punishing them. This is consistent with the result of a recent theoretical model predicting a mixed-strategy equilibrium where punishers and non-punishers co-exist—not everyone has to be a punisher (Boyd, Gintis, & Bowles, 2010). However, critics might question the external validity of this result because the laboratory version of third-party punishment (i.e., a lone punisher inflicts punishment against non-cooperators) is rarely observed in fields (Baumard, 2010; Guala, 2012). To reconcile the experimental results with the field observations, it is worth noting that people in fields often rely on a coordinated or centralized punitive measure to maintain large-scale cooperation (e.g., Rustagi et al., 2010; Wiessner, 2005). Although coordinated/centralized punishment is not the same as third-party punishment operationalized in laboratory (i.e., private

punishment), without punitive motivation, people might not even think about any centralized systems of punishment against non-cooperators. In addition, a recent study showed that although people preferred having the reward option to having the punishment option in the public goods experiment, punishment promoted cooperation more than reward (Sutter, Haigner, & Kocher, 2010). Therefore, we should not stop studying psychological mechanisms of punishment.

One limitation of this study is that we did not provide participants any explicit incentives to acquire a prosocial reputation. Jordan, Hoffman et al. (2016), in contrast, gave participants an explicit incentive to signal their prosociality to their potential interaction partners. This might appear to be a crucial flaw in studies that aim at testing the costly signaling hypothesis of third-party intervention strategies. Nevertheless, it is likely that prosocial behaviors, whose ultimate function is to obtain a good reputation, are at least partly driven by emotions (e.g., empathic concern). Accordingly, people may engage in prosocial behaviors (i.e., signaling behaviors) without any conscious awareness of their reputational effects because the emotional proximate mechanism commits people to adaptive behaviors (Frank, 1988). In future studies, it is desirable to investigate the effect of the presence of reputational incentives to closely examine the underlying (conscious and unconscious) motivations of these prosocial strategies. Further research on psychological mechanisms promoting different types of third-party intervention strategies seems necessary to fully understand large-scale cooperation among humans.

References

Almenberg, J., Dreber, A, Apicella, C. L., & Rand, D. G. (2011). Third party reward and punishment: Group size, efficiency and public goods. In N. M. Palmetti & J. P. Russo (Eds.), *Psychology of punishment* (pp. 73-92). New York: NOVA Science Publishers.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*, 325-344. doi:10.1016/j.evolhumbehav.2006.01.003

Batson, C. D. (1991). *The altruism question: Toward a social-psychological answer*. Hillsdale, NJ: Lawrence Erlbaum.

Batson, C. D. (2011). What's wrong with morality? *Emotion Review*, *3*, 230-236. doi:10.1177/1754073911402380

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, *40*, 290-302. doi:10.1037/0022-3514.40.2.290

Batson, C. D., Kennedy, C. L., Nord, L.-A., Stocks, E. L., Fleming, D. A., Marzette, C. M., Lishner, D. A., Hayes, R. E., Kolchinsky, L. M., & Zerger, T. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology*, *37*, 1272-1285. doi:10.1002/ejsp.434

Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind and Society*, *9*, 171-192. doi:10.1007/s11299-010-0079-9

Bone, J., Silva, A. S., & Raihani, N. J. (2014). Defectors, not norm violators, are punished by third-parties. *Biology Letters*, *10*, 20140388. doi:10.1098/rsbl.2014.0388

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*, 617-620.

doi:10.1126/science.1183665

Charness, G., Cobo-Reyes, R., & Jiménez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior and Organization, 68,* 18-28. doi:10.1016/j.jebo.2008.02.006

Clutton-Brock, T. H. & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373,* 209-216. doi:10.1038/373209a0

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for multidimensional approach. *Journal of Personality and Social Psychology, 44,* 113-126. doi:10.1037/0022-3514.44.1.113

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences, 8,* 185-190. doi:10.1016/j.tics.2004.02.007

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior, 25,* 63-87. doi:10.1016/S1090-5138(04)00005-4

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

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

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206,* 169-179. doi:10.1006/jtbi.2000.2111

Goetz, J. L., Keltner, D., & Simon-Thomas, E. (2000). Compassion: An evolutionary analysis and empirical review. *Psychological Bulletin, 136,* 351-374. doi:10.1037/a0018807

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35,* 1-59. doi:10.1017/S0140525X11000069

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology, 65*, 94-104. doi:10.1016/j.jesp.2016.04.004

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243-1248. doi:10.1126/science.162.3859.1243

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyantaz, A., Cardenas, J. C., Gruven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science, 327,* 1480-1484. doi:10.1126/science.1182238

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science, 312*, 1767-1770. dio:10.1126/science.1127333

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

Horita, Y. (2010). Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science, 1*, 6-9. doi:10.5178/lebs.2010.2

Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience,* 9:24. doi: 10.3389/fnbeh.2015.00024

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530,* 473-476. doi:10.1038/nature16981

Jordan, J., McAuliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics, 19*, 741-763. doi:10.1007/s10683-015-9466-8

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free-riding may be thwarted by second-order reward rather than punishment. *Journal of Personality and Social Psychology, 95*, 826-842. doi:10.1037/a0011381

Konishi, N., Oe, T., Shimizu, H., Tanaka, K., & Ohtsubo, Y. (in press). Perceived shared condemnation intensifies punitive moral emotions. *Scientific Reports*.

Konishi, N., & Ohtsubo, Y. (2015). Does dishonesty really invite third-party punishment? Results of a more stringent test. *Biology Letters*, 20150172. doi: 10.1098/rsbl.2015.0172

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75-84. doi:10.1016/j.evolhumbehav.2006.06.001

Leliveld, M. C., van Dijk, E., & van Beest, I. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology*, *42*, 135-140. doi: 10.1002/ejsp.87

Lotz, S., Baumert, A., Schlösser, T., Gresser, F., & Fetchenhauer, D. (2011). Individual differences in third-party interventions: How justice sensitivity shapes altruistic punishment. *Negotiation and Conflict Management Research, 4*, 297-313. doi:10.1111/j.1750-4716.2011.00084.x

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology*, *47*, 477-480. doi:10.1016/j.jesp.2010.10.004

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger,

J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B, 275*, 587-590. doi:10.1098/rspb.2007.1517

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1-10. doi:10.1016/j.cognition.2014.08.013

Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgement and Decision Making*, *4*, 543-553. http://journal.sjdm.org/91001/jdm91001.html

Ohtsubo, Y., Masuda, F., Watanabe, E., & Masuchi, A. (2010). Dishonesty invites costly third-party punishment. *Evolution and Human Behavior*, *31*, 259-264. doi:10.1016/j.evolhumbehav.2009.12.007

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*, 137-158. doi:10.1257/jep.14.3.137

Ozono, H., & Watabe, M. (2012). Reputational benefit of punishers: Comparison among punishers, rewarders and non-sanctionars. *Letters on Evolutionary Behavioral Science*, *3*, 21-24. doi:10.5178/lebs.2012.22

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans *really* punish altruistically? A closer look. *Proceedings of the Royal Society B, 280*, 20122723. doi:10.1098/rspb.2012.2723

Raihani, J. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology and Evolution*, *31*, 98-103. doi:10.1016/j.tree.2014.12.003

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, *69*, 993-1003. doi:10.1111/evo.12637

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology and Evolution, 27*, 288-295. http://dx.doi.org/10.1016/j.tree.2011.12.004

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in children. *Current Biology, 25*, 1731-1735. doi:10.1016/j.cub.2015.05.014

Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science, 330*, 961-965. doi:10.1126/science.1193649

Sutter, M., Haigner, S., & Kocher, M. G. (2010). Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies, 77*, 1540-1566. doi: 10.1111/j.1467-937X.2010.00608.x

Wiessner, P. (2005). Norm enforcement among the Ju/'hoansi bushmen: A case of strong reciprocity? *Human Nature, 16,* 115-145. doi:10.1007/s12110-005-1000-9

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*, 110-116. doi: 10.1037/0022-3514.51.1.110

Footnote

[1] We had anticipated that we would have a sufficient number of the strictly fair allocation (i.e., giving the partner 50 JPY), and use the strictly fair allocation as the control (fair) condition. However, there were an insufficient number of such strictly fair allocations made by the first set of participants. Accordingly, we included “giving 45 or 55 JPY to the partner” in fair allocations. We re-ran the analyses reported in the main text applying the strict fairness criterion, and found that the reported results would not be substantially altered by this strict criterion. The results of the re-analyses are reported in the Online Supplementary Analyses.

Table 1

Three Cross-Tables of the Number of Participants as a Function of Their Use of Two of the Three Intervention Options

Punishment* × *Help	Punished	Did Not Punish
Helped	18	30
Did Not Help	7	32
Punishment* × *Reward	Punished	Did Not Punish
Rewarded	19	35
Did Not Reward	6	27
Help* × *Reward	Helped	Did Not Help
Rewarded	45	9
Did Not Reward	3	30

Notes. The top panel shows the 2 × 2 cross-table based on participants' uses of the punishment and help options. The middle panel shows the 2 × 2 cross-table based on participants' uses of the punishment and reward options. The bottom panel shows the 2 × 2 cross-table based on participants' uses of the help and reward options.

Table 2

Emotional Correlates of Each of the Three Intervention Strategies

	Emotions Associated with the Unfair Allocator			Emotions Associated with the Recipient of the Unfair Allocation		
	Anger[†] (α = .90)	EC (α = .62)	Envy (α = .60)	Anger (α = .88)	EC (α = .78)	Envy (α = .24)
Punishment	.33** [.33**]	−.26* [−.26*]	−.02	−.15	.13	−.20^{+} [−.20^{+}]
Help	.21^{+} [.24*]	−.01	.03	−.17	.25* [.27*]	−.03
	Emotions Associated with the Generous Allocator			Emotions Associated with the Recipient of the Generous Allocation		
	Anger (α = .82)	EC (α = .78)	Envy (α = .25)	Anger (α = .81)	EC (α = .76)	Envy (α = .46)
Reward	−.16	.30** [.26*]	.22* [.19^{+}]	.01	.03	.22* [.17]

Notes. For each emotion score, Cronbach's alpha coefficient is reported in parentheses. Because of the low reliability associated with the envy scores, we avoided interpreting the results associated with the envy scores in this paper. The correlation coefficients in brackets are partial correlations that controlled for the effect of sex. "EC" designates "empathic concern."

[†] "Anger" in this cell qualifies as moral outrage.

** $p < .01$. * < $p < .05$. + $p < .10$.

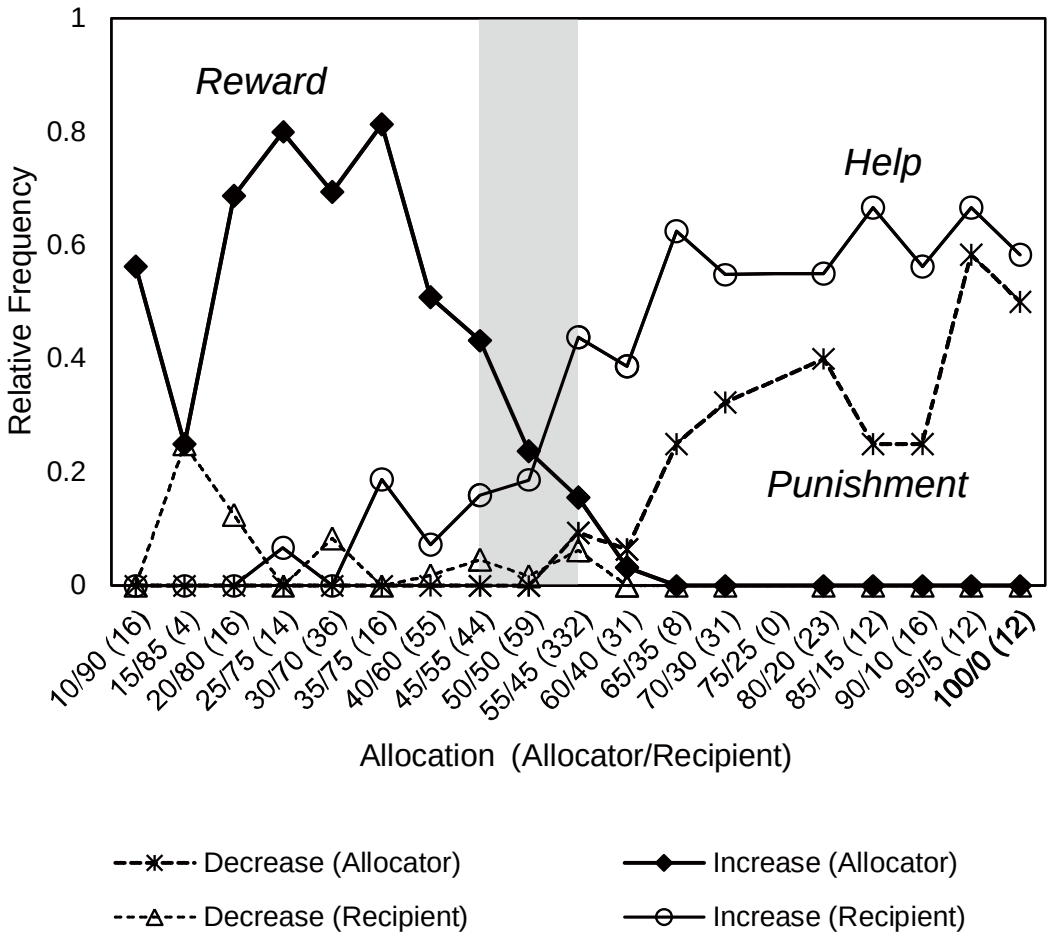

Figure 1. Relative frequencies of the use of each response option (i.e., decrease the allocator's payoff, increase the allocator's payoff, decrease the recipient's payoff, and increase the recipient's payoff) as a function of allocation. The gray area indicates the fair condition. The right-hand side of the gray area corresponds to the unfair condition. The left-hand side of the gray area corresponds to the generous condition.

Supplementary Analyses of "Within-Individual Associations among Third-Party Intervention Strategies: Third-Party Helpers, but Not Punishers, Reward Generosity"

Analyses with the Strict Criterion of the Fair Allocation

In the main text, we included the slightly unfair allocation (giving 45 JPY to the recipient and giving 55 JPY to the self) and the slightly generous allocation (giving 55 JPY to the recipient and giving 45 JPY to the self) in fair allocations. To exclude the possibility that the results reported in the main text are somehow dependent on this less strict definition of the fair allocation, we re-analyzed the data using the strict fairness criterion. In particular, we replaced 11 participants' responses to generous allocation (13% of the data) because they had seen the allocation of 55 JPY to the recipient before seeing more generous allocations. We also replaced 13 participants' responses to unfair allocation (15% of the data) because they had seen the allocation of giving 45 JPY to the recipient before seeing more unfair allocations. We re-ran most of the analyses reported in the main text using this new dataset.

Frequencies of Strategy Uses

We first confirmed whether participants used the three intervention strategies in relevant contexts. Since there were not a sufficient number of the strict fair allocations, there were only 59 participants whose five allocation games included the strictly fair allocation. In response to the first unfair allocation that they saw, 11 of 59 participants (.19) decided to reduce the unfair allocator's payoff, whereas no participants (.00) did so in response to the strictly fair allocation. We did not conduct McNemar's test because one of the 2 × 2 cells included 0 observation. For third-party help, 27 participants (.46) increased the recipient's payoff in response to the unfair

allocation, while 11 participants (.19) did so in response to the strictly fair allocation, $\chi^2(1) = 9.30$, $p = .002$. For third-party reward, 33 participants (.56) increased the allocator's payoff in response to the generous allocation, while 14 participants (.24) did so in response to the strictly fair allocation, $\chi^2(1) = 3.03$, $p = .082$. Although the result for third-party reward became only marginally significant (possibly due to the reduced sample size), these results were mostly consistent with the results reported in the main text.

We then tested whether participants were more likely to use prosocial intervention strategies than third-party punishment by conducting a series of McNemar's test. Confirming the results reported in the main text, participants were more likely to help the victim of the unfair allocator (.51) than punish the unfair allocator (.22), $\chi^2(1) = 15.36$, $p < .001$; they were also more likely to reward the generous allocator (.57) than to punish the unfair allocator (.22), $\chi^2(1) = 20.93$, $p < .001$; however, they were not significantly more likely to reward the generous allocator (.57) than to help the victim of the unfair allocator (.51), $\chi^2(1) = 1.25$, $p = .264$.

The sex difference pattern did not change from the original analyses. The sex difference in the punishment rate (.24 and .20 for men and women, respectively) was not significant by Fisher's exact test, $p = .796$. However, the sex difference was significant for the help rate (.27 and .68 for men and women, respectively), $p < .001$, and for the reward rate (.39 and .71 for men and women, respectively), $p = .004$ by Fisher's exact test.

Intra-individual Associations of Intervention Strategies

The analyses of the intra-individual associations of the intervention strategies mostly confirmed the original results (especially the most critical one). The intra-individual association between third-party punishment and help (Table S1) was not significant by Fisher's exact test, $p = .30$, which was marginally significant in the analysis reported in the main text. The

corresponding intra-individual correlation between punishment and help, *r*(punish, help), was .13. The intra-individual association between third-party punishment and reward (shown in the middle panel of Table S1, *r*(punish, reward) = .12) was not significant by Fisher's exact test, $p = .307$. For the intra-individual association between third-party help and third-party reward (shown in the bottom panel of Table S1, *r*(help, reward) = .54), it was significant by Fisher's exact test, $p < .001$.

Table S1

Three Cross-Tables of the Number of Participants as a Function of Their Use of Two of the Three Intervention Options

Punishment* × *Help	Punished	Did Not Punish
Helped	12	32
Did Not Help	7	36
Punishment* × *Reward	Punished	Did Not Punish
Rewarded	13	37
Did Not Reward	6	31
Help* × *Reward	Helped	Did Not Help
Rewarded	37	13
Did Not Reward	7	30

Notes. The top panel shows the 2 × 2 cross-table based on participants' uses of the punishment and help options. The middle panel shows the 2 × 2 cross-table based on participants' uses of the punishment and reward options. The bottom panel shows the 2 × 2 cross-table based on participants' uses of the help and reward options.

A series of Williams's tests for two correlated correlations indicated that r(help, reward) was significantly higher than r(punish, help), $t(84) = 3.31$, $p = .001$, and r(punish, reward), $t(84) = 3.47$, $p < .001$. Therefore, Hypotheses 1a and 1b from the main text were supported with the dataset produced by applying the strict fairness criterion.

Emotional Correlates of Each Strategy

We then computed point-biserial correlations among the three intervention strategies and emotional responses (see Table S2). Confirming Hypothesis 2a and the results in the main text, both moral outrage (i.e., anger at the unfair allocator) and reduced empathic concern for the unfair allocator were significantly correlated with third-party punishment. As shown in brackets, the two significant correlations supporting Hypothesis 2a remained significant after controlling for the effect of sex. However, Hypothesis 2b was not fully supported. Unlike the results reported in the main text, none of the emotional reactions were significantly correlated with third-party help. In particular, empathic concern for the recipient was only marginally significantly correlated with third-party help. However, as shown in brackets, once the effect of sex was controlled for, third-party help and empathic concern for the recipient of unfair allocation became significant.

In the generous condition, confirming Hypothesis 2c and the results reported in the main text, empathic concern for the generous allocator was significantly correlated with third-party reward. As shown in brackets, the partial correlation between empathic concern for the generous allocator and third-party reward became marginally significant after controlling for the effect of sex.

Table S2

Emotional Correlates of Each of the Three Intervention Strategies

	Emotions Associated with the Unfair Allocator			Emotions Associated with the Recipient of the Unfair Allocation		
	Anger[†]	EC	Envy	Anger	EC	Envy
Punishment	.33** [.33**]	−.24* [−.25*]	−.14	−.12	.11	−.17
Help	.17	−.04	−.02	−.19	.19[+] [.22*]	−.06

	Emotions Associated with the Generous Allocator			Emotions Associated with the Recipient of the Generous Allocation		
	Anger	EC	Envy	Anger	EC	Envy
Reward	−.15	.25* [.21[+]]	.19[+]	.02	.03	.19[+]

Notes. For each emotion score, Cronbach's alpha coefficient is reported in parentheses. Because of the low reliability associated with the envy scores, we avoided interpreting the results associated with the envy scores in this paper. The correlation coefficients in brackets are partial correlations that controlled for the effect of sex. "EC" designates "empathic concern."

[†] "Anger" in this cell qualifies as moral outrage.

** $p < .01$. * $< p < .05$. + $p < .10$.