Altruistic Norm Enforcement and Decision-Making Format in a Dilemma: Experimental Evidence

Kenju Kamei

Durham University

6 February 2017

Altruistic Norm Enforcement and Decision-Making Format in a Dilemma: Experimental Evidence

Kenju Kamei[#]

Department of Economics and Finance, Durham University
Email: kenju.kamei@gmail.com, kenju.kamei@durham.ac.uk

This version: February 2017

Abstract:

Past research has shown that people often take punitive actions towards norm violators even when they are not directly involved in transactions. However, it at the same time suggests that such third-party punishment may not be strong enough to enforce cooperation norms in dilemma situations. This paper experimentally compares the effectiveness of third-party punishment between different enforcement formats. Consistent with past studies, our data shows that having an individual third-party punisher in a group does not make one's defection materially unbeneficial because of the weak punishment intensity. It also shows that third-party punishment is not effective when two individuals form a pair as a punisher and jointly decide how strong third-party punishment they impose. However, third-party punishment can be sufficiently strong to enforce cooperation norms when a third-party punisher's action choice is made known to another individual third-party punisher in a different group, or when there are two independent individual third-party players in a group.

JEL classification: C92, D01, H49
Keywords: experiment, cooperation, dilemma, third-party punishment, social norms

[#] We thank John Hey for his hospitality when we conducted experiments in the University of York. We also thank Mark Wilson (an IT manager at the University of York) for his support in setting the computers in the experimental sessions. We also thank Lan Hoang Dinh (a former master student at Bowling Green State University) for his research assistance in ztree programming. This project was funded by Durham University Business School, and the Kyoto University Institute of Economic Research foundation.

1. Introduction

One of the most commonly observed features of people's interactions is cooperation dilemmas, such as prisoner's and collective action dilemmas. A cooperation dilemma is a situation where agents receive benefits from peers' cooperative actions but no one takes cooperative actions if they are purely selfish. For the last few decades, scholars have intensively studied people's cooperation behaviors and mechanisms that may sustain cooperative social norms in societies, both theoretically and empirically (e.g., Chaudhuri 2011, Fehr and Fischbacher 2004b, Fehr and Schmidt 2006, Van Lange *et al.* 2011 for a survey).

One well-known behavioral principle in the recent few decades is that people have other-regarding preferences, such as inequity aversion (e.g., Fehr and Schmidt 2006 and Sobel 2005 for a survey). Initiated by the seminal work by Fehr and Fischbacher (2004a), a large body of experimental research has shown that even third parties, who are not directly involved in two-party interactions, frequently impose punishment when they encounter unfair economic interactions in dilemma games (e.g., Fehr and Fischbacher 2004a, Carpenter and Matthews 2012, Kurzban *et al.* 2007, Lergetporer *et al.* 2014). Many scholars have uncovered a number of features of such third-party punishment, especially by using dictator games with third-party punishment. For instance, it is now known that altruistic third-party punishment is associated with activation of some brain regions, such as reward regions (e.g., Buckholtz *et al.* 2008, Strobel *et al.* 2011 Corradi-Dell'Acqua *et al.* 2012). Past research also shows that the intensity of third-party punishment differs by various factor – individual characteristics, such as justice sensitivity (e.g., Lotz *et al.* 2011), population and the size of societies (e.g., Henrich *et al.* 2006, Marlowe *et al.* 2010), group affiliations, i.e., whether norm violators belong to the punisher's

group (e.g., Bernhard *et al.* 2006, Lieberman and Linke 2007), and so forth. The third-party punishment behavior is considered to be unique to humans (e.g., Riedl *et al.* 2012).

Although individuals' third-party punishment is often observed, it is by far weaker than second-party punishment in most past research. For instance, Fehr and Fischbacher (2004a) conclude that "sanctions by second parties directly harmed were much stronger than third-party sanctions, indeed strong enough to make norm violations unprofitable, whereas the sanctions of a single third party were not" (page 85).[1] The weak intensity of the third-party punishment has been replicated by subsequent studies (e.g., Carpenter and Matthews 2012, Lergetporer *et al.* 2014).[2] Even when subjects endogenously choose to act as third-party punishers, the level of cooperation declines steadily over time, because the endogenous choice oppositely decreases the intensity of third-party punishment although defectors' responses to punishment increases (Marcin *et al.* 2016). However, in our real life, despite its weak strength seen at the past experimental studies, researchers argue that third-party punishment may substantially help sustain cooperation norms, especially in large-scale societies (e.g., Henrich *et al.* 2006, Marlowe *et al.* 2010). How can we reconcile the seemingly conflicting observations of the important role of third-party punishment in reality and its weak effect seen in the past experimental research? To the best of our knowledge, with what mechanisms third-party punishment become deterrent is one important remaining question.

This paper experimentally explores how the strength of third-party punishment differs by the third party's decision-making format. This study consists of five treatments. In each

[1] The literature also suggests that providing third parties an opportunity to reward second-party players likewise does not help cooperation norms to evolve in a dilemma situation (e.g., Sutter *et al.* 2009)

[2] The experiment by Lergetporer *et al.* (2014), using children (whose ages were 7 to 11 years old), showed that subjects incorrectly anticipated that third-party punishment was very common and thus believed cooperation would be materially more beneficial, although defection was in fact still the payoff-maximizing action considering the actual frequency of punishment. See McAuliffe *et al.* (2011) for evidence on the prevalence of third-party punishment by young children in a dictator game.

treatment, there are two subjects in a group that play a one-shot prisoner's dilemma game with each other. In Fehr and Fischbacher (2004a), stand-alone third-party individuals are placed along with the two-party players and then the third-party individuals decide how to impose sanctions to the two-party players whom they are assigned. We use this setup as the control treatment. We then set up four treatment conditions by varying the decision-making formats of third-party punishers. In the first treatment condition, two third-party punishers from different groups are paired, but independently make third-party punishment decisions in their own groups and then each punisher's action choice is made known to the paired punisher in the enforcement team. The design of this treatment is motivated by the literature which suggests that the visibility of acts may affect people's pro-social behavior (e.g., Sell and Wilson 1991, Kamei and Putterman 2015). There is also real-world relevance to this treatment. For instance individuals who work in public enforcement usually share reports with others who work in the same area or division if they encountered with law violators.[3] The second treatment condition involves changes of the format to inflict punishment. Instead of simply making two third-party individuals paired while having a transparent decision process, we let a pair of two third-party individuals face with the same two-party interaction and then let them *jointly* make a single punishment decision to the two-party players. We note that each third-party pair is anonymously and randomly formed at the onset of the experiment by the computer. The punishment behavior of third-party pairs may be different from that of third-party individuals because a large volume of the literature suggests that decision-making in teams may affect people's behavior (e.g., Charness and Sutter 2012, Kugler *et al.* 2012, Kerr and Tindale 2004 for surveys). In the third treatment condition, two third-party individual punishers are placed in a group with two-party players and they

[3] This information sharing is automatic and does not involve subjects' choices in the experiment in order to simplify the experiment.

independently and simultaneously make punishment decisions. The action choices of the two third-party punishers are made open to each other as in the first treatment condition. The effectiveness of having two individual punishers is unclear because one punisher may attempt to free ride on the other punisher. Lastly, in the fourth treatment, whose overall design is identical to that in the above second treatment, pairs are formed while revealing their identity within pairs. The fourth treatment condition is designed because the degree of social distance within pairs may affect pairs' pro-social behavior (see Cason and Mui (1997) for a dictator game; Kamei (2016) for a public goods game; and Kocher and Sutter (2012) for a gift-exchange game). In short, this paper contributes to a rich body of the literature on altruistic punishment by providing new evidence on how the difference in the decision-making format of third-party players affects people's altruistic punishment behavior.

Our result on the target of punishment replicates the finding of Fehr and Fischbacher (2004a) and other related studies. In each of the five treatments, third-party punishers impose significantly stronger punishment on a defector who is faced with a cooperator than any other type of second-party player. However, the intensity of third-party punishment differs by treatment. First, in the control treatment, although such third-party punishment is frequently observed, it is so weak that choosing to defect is still materially more beneficial for a second-party player, as is consistent with the past research. However, this does not hold in the first treatment condition. Once each third-party punisher is paired with another punisher and their action choice is made known to another punisher, the strength of third-party punishment becomes more than double. The impact of raising visibility of punishment actions is strong enough to make choosing to cooperate materially more beneficial for second-party players. But, intriguingly, once two-person third-party punishment pairs are faced with the same second-party

players and are asked to jointly make a single punishment decision towards them, the strength of third-party punishment becomes similar to that of the control treatment. A close look at the data also indicates that prior to communicating with the other paired punishers, third-party pairs have a similar level of willingness to punish to individual punishers in the treatment where their punishment acts are visible to other punishers (the first treatment condition already explained). However, the pairs decrease their willingness to punish significantly through the joint decision-making process (communication). Similar results are found in the joint-punishment treatment where the social distance between the two punishers is small. Lastly, the strength of punishment is clearly different when two independent third-party players, instead of a single third-party punisher, are grouped with two-party players. The data shows that although we observe free-riding behavior among two individual third-party players, the presence of two punishers is deterrent enough to make the second-party players choose cooperation.

The rest of the paper proceeds as follows: Section 2 describes our experimental design. Section 3 briefly provides hypotheses along with related literature. Section 4 reports results, and Section 5 concludes.

2. Experimental Design

The design frame of our study is a prisoner's dilemma game with a third-party punisher (Fehr and Fischbacher 2004a). We use a between-subjects design. This means that each subject plays the game only with one treatment condition. There are five treatments which vary in the third-party punishment form: whether the third party is (a) an independent individual, (b) an individual but has another individual as an enforcement team (they act individually in different groups but each punisher's action choice is informed to the other individual in the enforcement team), (c) a pair of individuals that jointly decide a single punishment amount in their group as

an enforcement team, or (d) two third-party individuals that independently make punishment decisions toward the same targets in a group. As for (c), we further design two treatments by varying the social distance between paired individuals (see Section 2.3 for the detail). We call the five treatments the "Individual Punishment" treatment (abbreviated as the I-P treatment), the "Individual Punishment with Partner" treatment (abbreviated as the I-P-P treatment), the "Joint Punishment" treatment (abbreviated as the J-P treatment), the "Joint Punishment, Close Social Distance" treatment (abbreviated as the J-P-C treatment), and the "2 Individual Punishment towards the Same Target" treatment (abbreviated as the 2-I-P-S treatment). The conversion rate in the experiment is as follows: five points in the experiments are equal to £1 in all treatments. Each treatment consists of two stages. In the first stage, individuals who are assigned the role of two-party players play one-shot prisoner's dilemma games (Fehr and Fischbacher 2004a). Specifically, two individuals are each given an endowment of 25 points and simultaneously decide whether or not to send 10 points to their counterparts. If a subject sends 10 points to her counterpart, the 10 points are tripled and become a payoff of the recipient; and the remaining 15 points become the sender's payoff. If the subject does not send 10 points to the other player, the 25 points become the payoff of her own. In each group (interaction unit), there is another individual(s) or a pair of individuals who is not involved in the dilemma game. Those third-party players are asked to answer their guesses on how many persons in the group (= {0, 1, 2}) they think will send 10 points to the counterparts.[4]

In the second stage, third-party players decide how many punishment points they want to assign to each of the two-party players in their groups, contingent on the two players' action

[4] As our focus is on third-party players' action choices, this question is not incentivized to avoid some subsequent effects on their punishing behaviors in the next stage. See Gächter and Renner (2010) for possible effects of having incentivized elicitation of beliefs. We note that this question is included to keep high anonymity in sessions: with this additional question, the numbers of computer mouse clicks are the same between players that play the dilemma game and third-party players.

choices (send or not send) in stage 1 (the third-party punishers make decisions before being informed of the first stage outcome). The punishers are given an endowment of 40 points each. Punishment points assigned to each two-party player must be an integer between 0 and 20. The cost ratio of the punishment technology (the punished: the third-party individual) is 3:1. That is, for each punishment point assigned to a target, one point is deducted from the third-party punisher's payoff and three points are deducted from the target's payoff. We employ a strategy method for the third-party punisher's decisions in order to obtain as many incentive-compatible observations as possible. Specifically, each third-party punisher is asked to answer reduction points they would assign under the following four possible scenarios in stage 1:

(a) Reduction points targeted to a player that sent 10 points to his or her counterpart when the counterpart also sent 10 points to that player;

(b) Reduction points targeted to a player that did not send 10 points to his or her counterpart when the counterpart sent 10 points to that player;

(c) Reduction points targeted to a player that sent 10 points to his or her counterpart when the counterpart did not send 10 points to that player;

(d) Reduction points targeted to a player that did not send 10 points to his or her counterpart when the counterpart also did not send 10 points to that player.

After third-party punishers make decisions, their choices in one of the four scenarios will be applied dependent on realized interaction outcomes in Stage 1. In case that the payoff of a second-party player is negative due to punishment received, the player's payoff is set zero. While third-party players are deciding on punishment points, two-party players who played the prisoners dilemma game in stage 1 are asked to submit their guesses on how many punishment

points they expect to receive if each of the four possible scenarios mentioned above happened in their groups.[5] Once stage 2 is over, each subject learns the outcome of both stage 1 and stage 2.

In the experiment, subjects' identity remains anonymous in order to measure the impact of one's punishment decisions being informed to another third-party punisher and that of the joint decision-making protocol in a controlled manner. However, we relax this setup in the J-P-C treatment by letting paired third-party punishers know each other's identities in the joint decision-making process to study the robustness of the findings from the J-P treatment.

We now explain each treatment one by one.

2-1. The I-P treatment

In the I-P treatment, subjects are randomly assigned to a group of three at the beginning of the experiment. Three players in each group are randomly (i.e., with a probability of 1/3) assigned a player number: either player 1, 2 or 3 so that each player has a different number in the group. Subjects who are assigned the roles of player 1 and player 2 play the aforementioned one-shot prisoner's dilemma game in the first stage. Subjects who are assigned the role of player 3 act as the third-party punishers. Player 3s decide how many punishment points they want to assign to player 1s and 2s under the strategy method explained above.

2.2. The I-P-P treatment

At the onset of the I-P-P treatment, subjects are randomly assigned a player number: either player 1, 2, 3 or 4 so that one-third of the subjects are player 1, one-third of them are player 2, one-sixth of them are player 3 and the remaining one-sixth are player 4. Each player 3 is then randomly and anonymously paired with a player 4 as an enforcement team. All subjects

[5] As in some other hypothetical questions, this questionnaire is included for the purpose of keeping high anonymity in sessions: with this additional question, the numbers of computer mouse clicks are four and are the same between second-party players (player 1s and 2s) and third-party players (player 3s and 4s).

are stochastically assigned to a group of three so that (a) there are always one player 1 and one player 2 in each group and (b) there is either player 3 or player 4 in each group. Even though a player 3 and a player 4 are put in a team, they belong to different groups as explained, and act individually to decide how many punishment points to give toward different targets.

Stages 1 and 2 proceed exactly the same as the I-P treatment, except that the punishment behaviors of player 3s and 4s are informed to the paired punishers after Stage 2. Each punisher is not informed of their pair partner's third-party punishment points under scenarios that were not applied. We can measure the impact of having a pair where a subject's third-party behavior is made known to the third-party partner, which we call the "social effects" in the paper, by comparing the intensity of third-party punishment between the I-P and I-P-P treatments.

2.3. The J-P treatment and the J-P-C treatment

In the J-P treatment, subjects are randomly assigned to a group of four. Four players in each group are randomly (i.e., with a probability of 1/4) assigned a player number: either player 1, 2, 3 or 4 so that each player has a different number in the group. Player 3s and 4s then form pairs in their groups, and the pairs act as the third-party punishers.

Stage 1 proceeds the same as the I-P and I-P-P treatments. Once player 1s and 2s decide how much to send, player 3s and 4s are given five minutes to freely discuss with their pair partners via a computer chat window, as is similar to Kamei (2016).[6] Subjects are neither informed at which desk their counterparts are seated in the experimental laboratory nor given any other information that may specify the matched third-party partners' identity. Before communicating with partners, each third-party individual in a pair is asked to answer how many

[6] In Kamei (2016), one minute is given to paired subjects in order to jointly make one contribution decision in each period of a repeated public goods game. A longer duration, five minutes, is given to subjects in the present study. This is done so because each pair needs to make four joint decisions unlike Kamei (2016).

punishment points they would assign to each of player 1 and 2 as the pair's joint punishment points under each of the four scenarios if they could decide their pair's joint punishment amounts unilaterally without communicating with partners (on condition that the payoff consequence would be the same between the two individuals in the pair as that in the experiment). This elicitation task is included to investigate possible social effects in the J-P treatment as an additional analysis.[7] If having a pair-mate inflates punishers' non-material motives, such as inequity aversion, third-party punishers may exhibit inflated willingness to punish before the communication begins with their partners. While player 3s and 4s are answering the questionnaire, player 1s and 2s are asked about their expectations concerning how many punishment points they would receive under the four possible scenarios.[8] While third parties (player 3s or 4s) are communicating with each other, player 1s and 2 are asked to answer open-ended questions on their computer terminals.[9]

Once the five minutes of the discussion stage passes, each individual in pairs submits punishment points they want to assign as a pair under the four possible scenarios. The joint decision-making rule is as follows: if two individuals in a pair submit the same punishment points in a given scenario, then the points become their pair's joint punishment points. If they submit different points, then one of the two is selected with a probability of 50% by the computer as the pair's joint punishment points.[10] There are no real decisions to make for both player 1s and

[7] An alternative to eliciting hypothetical willingness to punish is to make the pre-communication question incentive-compatible by setting it to be realized with some probability. We did not employ this method because this way makes the design more complex to subjects and also because our focus is pairs' action choices. This hypothetical question can be used as supplementary evidence of social effects which we formally study with the I-P-P treatment.
[8] This questionnaire is not incentivized. It is just included so that the numbers of mouse clicks are the same among all players in order to preserve a high degree of anonymity in sessions.
[9] This task is also included to make all subjects equally occupied during sessions, regardless of assigned task.
[10] This rule was applied for only 2 pairs in Scenario (a), 4 pairs in Scenario (b), 3 pairs in Scenario (c), and 3 pairs in Scenario (d) in the J-P treatment. This rule was applied for only 1 pair in Scenario (a), 2 pairs in Scenario (b), 1 pair in Scenario (c), and 2 pairs in Scenario (d) in the J-P-C treatment.

2s during this stage.[11] Once all third-party punishers submit punishment points for the four scenarios, they are informed of realized interaction outcomes.

The J-P-C treatment is designed to supplement the J-P treatment so that we can investigate how social distance within pairs may alter the punishment behavior of individuals in pairs. The J-P-C treatment is identical to the J-P treatment, except that each subject in a pair is informed of the pair partner's seat number and is then given two minutes to introduce themselves before the five-minute discussion about their joint punishment acts. There are no restrictions in the contents of communication except that any offensive languages are prohibited; and subjects are explained that communication contents are private, are not subject to analysis, and are not disclosed in any formats. Thus, social distance within pairs is smaller in the J-P-C treatment than in the J-P treatment.

2.4. The 2-I-P-S treatment

In the 2-I-P-S treatment, as in the J-P and J-P-C treatments, subjects are randomly assigned to a group of four and are given a player number: either player 1, 2, 3 or 4 so that each player has a different number in the group. In the 2-I-P-S treatment, there are two independent, individual third-party punishers (player 3 and player 4) along with player 1 and player 2 in each group, instead of a single individual punisher in the group in the I-P and I-P-P treatments. All of the other design pieces, such as the interactions between player 1 and 2 in stage 1 and the punishment technology (the cost ratio of 1:3), are the same as the other treatments. As in the I-P-P, J-P and J-P-C treatments, each third-party punisher is informed of the other punisher's punishment points imposed in the group at the end of stage 2. Notice that the 2-I-P-S treatment is

[11] Player 1 and 2 are asked to answer hypothetical questions as to how many third-party punishment points they would impose if they observed the outcome of stage 1 which was identical to their own interaction outcome. These hypothetical questions were included by the same reason explained in footnote 9.

identical to the I-P-P treatment, except that the two punishers independently assign punishment points toward the *same* two individuals in their group in the 2-I-P-S treatment. Thus, the difference in the per-subject punishment intensity between the 2-I-P-S and I-P-P treatments can be attributed to the effects of subjects' intention to free ride on their third-party peers.

Finally, we note that for third-party punishment to work effectively, Fehr and Fischbacher (2004a) argue that: "in the context of our experiment, more than one third party is needed to enforce the norm." (page 85). The 2-I-P-S treatment also serves as a test for this explanation by Fehr and Fischbacher (2004a).

3. Related Literature and Discussions

This paper's focus is on the third party's punishment behavior. The standard theory prediction for the third party's behavior is straightforward in all treatments. It predicts that no third-party players inflict punishment points on second-party players because they do not interacts with second-party players in the experiment and the punishment acts are costly. However, some outcome-based other-regarding preference models, such as Fehr and Schmidt (1999), predict positive punishment behavior by third-party players in our environment (Fehr and Fischbacher 2004a).[12] As mentioned earlier, it is well-known that although third-party punishment is widespread in Scenario (b), the strength is not strong enough to make selecting to cooperate materially more beneficial for second-party players in such a dilemma situation (e.g., Fehr and Fischbacher 2004a, Lergetporer *et al.* 2014).[13] As the I-P treatment is based on Fehr and Fischbacher (2004a), we expect a similar result from this control treatment.

[12] Pure direct reciprocity-based models, such as Rabin (1993), do not explain third parties' punishment behavior because second-party players do not have any interactions with the third parties. Emotions, such as anger, may partly account for third party punishment (e.g., Nelissen and Zeelenberg 2009)

[13] We nevertheless note that there is some past research on other games which suggests that the intensity of third-party punishment may not be different from that of second-party punishment (e.g., Leibbrandt and López-Pérez 2012 for the context of a dictator game).

Hypothesis 1: *Third-party punishment in Scenario (b) is common, but the strength is such a weak level that selecting to defect is still the materially beneficial action for second-party players in the I-P treatment.*

However, changing the third-party punishment procedure may affect the third party's behavior. The closest treatment to the I-P treatment is the I-P-P treatment. In the I-P-P treatment, each third-party punisher's realized decisions are informed to their enforcement team-mate that acts in a different group. The literature suggests that increasing visibility on people's actions may affect their altruistic behavior in various contexts. For instance, Sell and Wilson (1991), in a finitely-repeated public goods game, found that subjects contribute more when individualized information on group members' contribution amounts is provided than otherwise.[14] Kamei and Putterman (2015) found that subjects achieve high efficiency with a punishment opportunity when individualized information on contributions and an opportunity to punish anyone are provided. In addition, being informed of others' pro-social behavior when choosing own actions is known to enhance people's pro-social behavior. For instance, Shang and Croson (2009), in a field experiment, indicated that radio listeners donated higher amounts in the station's on-air fund drive when they were provided social information that another member (without revealing the member's identity) contributed a high amount. Kamei (2014) found that subjects' second-party punishment points are positively related to other members' punishment toward a target. These kinds of results may have been partly caused by high visibility of action choices, which triggers social effects, such as shame, guilt and pride (e.g., Bowles and Gintis 2005), and

[14] We note that under which exact condition individualized information helps enhance cooperation norms is still unsettled, according to our knowledge. Some studies found different results from Sell and Wilson (1991). For instance, Weimann (1994) found that such individualized information has little impact on subjects' contribution behavior, and Wilson and Sell (1997) found it has instead negative impact on subjects' contribution behavior.

potentially influencing the third-party punishment behavior.[15] Further, there is evidence on the positive impact of social image on subjects' behavior in a number of contexts. For instance, Ariely *et al.* (2009) found that subjects work harder for charity in a real-effort experiment where they had to tell others about the amounts donated than when their identities and acts remain anonymous, when there are no private incentives associated with the effort task. Samek and Sheremeta (2014) found that subjects contribute significantly more when identifiable information of all group members or identifiable information of the lowest contributors, along with the members' contribution amounts, are revealed in each group, compared with the baseline case without identifiable information. In the context of third-party punishment, Kurzban *et al.* (2007) find that the frequency of punishment events and its strength are significantly larger when punishers are informed that their action choices are made known to the experimenter and will meet with him or her outside the laboratory than otherwise. They also find that subjects impose even larger punishment more frequently if a third-party individual's punishment behavior is known by all participants in a session while her own identity is also revealed. In the I-P-P treatment, each third-party subject's punishment behavior is informed to their enforcement partner without their identity being disclosed. However, even with punishment behavior revealed to another punisher only, subjects' altruistic punishment behavior may be enhanced due to some social effects.

Hypothesis 2: *Third-party punishment in Scenario (b) is more frequent and stronger in the I-P-P treatment than in the I-P treatment.*

How do joint punishment opportunities affect people's punishment behaviors? First, having another member in the enforcement team as in the I-P-P treatment may enhance pairs'

[15] Sell and Wilson (1991) argued that less free riding may occur because subjects could employ trigger strategies based on the information. We also note that the impact of social information, such as one seen in Shang and Croson (2009), can also be explained by people's conditional cooperation behavior.

willingness to punishment significantly as similar to Hypothesis 2. However, at the same time, there is a possibility that pairs decrease willingness to punish for at least two reasons; and we therefore cannot have a clear hypothesis. First, recent literature on institutions suggests that collective choices of institutions by majority voting may limit anti-social and irrational choices in public goods dilemmas (e.g., Ertan *et al.* 2009, Putterman *et al.* 2011, Kamei *et al.* 2015).[16] Because third-party punishment acts in all the four scenarios decrease the third party's payoff, the joint decision-making procedure may undermine pairs' willingness to punish. Second, a large body of the literature suggests that a joint decision-making process may make people more rational in a number of settings, including ultimatum games, beauty-contest games, signalling games, and centipede games (see, for example, Charness and Sutter (2012) and Kugler *et al.* (2012) for a survey), although they could behave more cooperatively if reputation concerns are strong in repeated setups (e.g., Gillet *et al.* 2009, Kamei 2016, Müller and Tan 2013). If two-person pairs behave more like a game theorist in our experiment, pairs would decrease their willingness to punish through communication.[17]

There are two individually acting third-party punishers in the 2-I-P-S treatment. There are possible opposing hypotheses for the effectiveness of having two individual third-party punishers. On the one hand, social effects potentially present in the I-P-P treatment may make individual punishment acts stronger in the 2-I-P-S treatment, compared with the I-P treatment. If this is the case, since the number of punishers are two, total punishment targeted at a defector could become strong enough to deter the two-party players' opportunistic behaviors. But on the other

[16] Also see van Miltenburg *et al.* (2014).

[17] A related paper to the J-P and J-P-C treatment is Auerswald *et al.* (2016), in which three-person teams repeatedly interact with other three teams in public goods game with a random matching and punishment decisions are made jointly towards interaction partners by voting in some treatments. Our paper is sufficiently different from Auerswald *et al.* (2016) for two important aspects. First and most importantly, two-person pairs in our paper engage in *third*-party punishment acts while three-person pairs Auerswald *et al.* engage in *two*-party punishment. Second, pairs in our paper make joint decisions through communication while teams in Auerswald *et al.* do so via voting.

hand, a punisher may attempt to free ride on the pair partner's third-party punishment acts considering that two third-party enforcers act independently in the group. If the negative effect of free-riding is large, altruistic punishment in this treatment may not be sufficiently strong to make defection material unbeneficial even though there are two punishers.[18]

4. Results

We conducted 15 sessions at the Centre for Experimental Economics laboratory (EXEC laboratory) at the University of York in December 2015 and February, September and October 2016.[19] A total of 280 students – 48 students for the I-P treatment, 48 students for the I-P-P treatment, 72 students for the J-P treatment, 68 students for the J-P-C treatment, and 44 students for the 2-I-P-S treatment – participated in the experiment. All subjects were recruited through Hroot (Hamburg registration and organization online tool; see Bock 2014). Solicitation messages were sent to all eligible subjects who have registered in the database and subjects voluntarily signed up for and participated in the sessions. No subjects participated in more than one session. All experiments except instructions were programmed in ztree (Fischbacher 2007). The instructions (available in online Appendix B) and verbal explanations in the experiment were neutrally framed. Any loaded words, such as cooperate and punish, were avoided. Subjects had

[18] The idea of free riding on others' altruistic punishment has been recently studied in a dictator game (Lewisch *et al.* 2011); which showed that a third-party player inflicts lower punishment points to a dictator that sends a small portion to the recipient, when there is another punisher in his/her group. Lewisch *et al.* (2011), nevertheless, found that the aggregate punishment points imposed on selfish dictators are larger than the punishment points when there is only one third-party punisher in a group. Carpenter and Matthews (2012) is the closest paper to the 2-I-P-S treatment. They let subjects impose third-party punishment (punishing subjects in other groups) along with second-party punishment in a finitely-repeated public goods game whose group size is four. Carpenter and Matthews (2012) found that although people spend more resources on second-party punishment, rather than third-party punishment, as in Fehr and Fischbacher (2004a), the frequency of third-party punishment is much lower compared with Fehr and Fischbacher (2004a). This may mean that subjects free rode on others' altruistic punishment acts in Carpenter and Matthews (2012). The 2-I-P-S treatment in our study is sufficiently different from Carpenter and Matthews (2012), as both second-party and third-party punishment opportunities were simultaneously given to the subjects in Carpenter and Matthews (2012). Thus, we cannot conclude whether subjects free ride on other peers' third-party punishment only from the data of Carpenter and Matthews (2012).

[19] We set the number of subjects in each of the J-P and J-P-C treatments around 1.5 times more than the other treatments. This was done so because we wanted to analyze third-party pairs' punishment decisions in details (e.g., the effects of communication) as will be explained in Section 4.4.

to answer a number of control questions included in the instructions and the experimenter explained the answers using whiteboards in order to make sure that the subjects understood the experiments fully. We first overview the sending decisions of player 1s and 2s in Section 4.1. We then explore the third-party punishment behaviors in Sections 4.2 to 4.4.

4.1. Sending Decisions of Second-Party Players

We find that around 66% of player 1s and 2s sent 10 points to their counterparts in the experiment. The sending rates differ slightly by treatment. They are 71.9% (23 out of 32 subjects), 71.9% (23 out of 32 subjects), 55.6% (20 out of 36 subjects), 70.6% (24 out of 34 subjects) and 54.6% (12 out of 22 subjects) in the I-P, I-P-P, J-P, J-P-C and 2-I-P-S treatments, respectively. The sending rates are not significantly different between any two treatments according to two-sample Fisher's exact tests (Part (1), Appendix Table A.1).

The distributions of third parties' expectations as to the number of persons who would send 10 points are very similar among the five treatments. They on average believed that 1.31, 1.44, 1.33, 1.35 and 1.27 persons would send 10 points in the I-P, I-P-P, J-P, J-P-C and 2-I-P-S treatments, respectively. Mann-Whitney tests fail to reject the null hypothesis that the average beliefs are different between any two treatments (Part (2), Appendix Table A.1).

RESULT 1: *Sending decisions of two-party players (player 1s and 2s) do not differ by format of third-party punisher. Third-party punishers' beliefs on the number of cooperators do not differ by format of third-party punisher.*

4.2. Punishment Decisions of Third-Party Players

Figure 1 shows the frequencies of third-party punishment and the average punishment points received by second-party players. Our data replicate those of past research. First, third-party punishment is common and the frequency is much higher in Scenario (b) – when a second-

party player defects but the counterpart cooperates, compared with any other scenarios (Panel (i)).[20] Second, second-party players receive much more punishment points from the third-party punishers in Scenario (b) (Panel (ii)).[21] This result is consistent with Fehr and Fischbacher (2004a). This punishment pattern resonates with the idea that people are inequality-averse agents and third-party players attempt to mitigate income inequality in their groups by engaging in third-party punishment. Third, we find that the punishment intensity in Scenario (d) is stronger than that in Scenario (a); and the difference between them is significant (weakly significant) in the I-P-P treatment (the J-P, J-P-C and 2-I-P-S treatments). This cannot be explained by subjects' inequality-averse preferences. This pattern is also found in Fehr and Fischbacher (2004a); who interpret that this behavior is consistent with the model by Levine (1998).[22]

RESULT 2: *Third-party punishment is common, especially in Scenario (b). Second-party players receive stronger punishment in Scenario (b) than in any other scenarios, as is consistent with past research.*

A comparison across the treatments finds that the punishment intensity imposed on a defector in Scenario (b) differs by decision-making format (Appendix Table A.4).[23] The average punishment received by defectors in Scenario (b) is 3.00 points in the I-P treatment. This is in sharp contrast with the I-P-P treatment where each punisher's action choice is made known to another person in her enforcement team: it is 6.63 points in the I-P-P treatment. The difference in the punishment intensity is significant between the two treatments (two-sided Mann-Whitney test,

[20] According to a two-sided Fisher's exact test based on the data of all treatments, the frequency of third-party punishment in Scenario (b) is significantly higher than that in the other scenarios (Appendix Table A.2).
[21] The punishment intensity in Scenario (b) is significantly stronger than that in Scenarios (a) and (c) in all of the five treatments. The former is also significantly stronger than that in Scenario (d) in the I-P-P, J-P and 2-I-P-S treatments (See Appendix Table A.3).
[22] Levine (1998) has proposed that a subject's utility is a linear function of her own payoffs and others' payoffs (the subject's utility could be negatively dependent on others' payoffs).
[23] Individual-level data (for the I-P treatment) and pair-average data (for the other four treatments) are used for comparing punishing intensity of the third party between the treatments in this section.

p = .0182). The structure of the J-P treatment is the same as that of the I-P-P treatment, except the following two aspects: (i) two persons in a team are faced with the same second-party players and (ii) the two persons jointly decide on single third-party punishment points through communication in the J-P treatment. With the joint decision-making protocol, the punishment intensity decreases to a much lower level, compared with the I-P-P treatment. The average punishment points the third-party players assigned in Scenario (b) is 3.31 points in the J-P treatment. This level is significantly different from the I-P-P treatment at the 10% level (two-sided Mann-Whitney test, p = .0725). This result does not depend on how pairs are formed. The punishment intensity in Scenario (b) in the J-P-C treatment (where identity of each partner is revealed within their pair in the pairing process) is 2.88 points and is not significantly different from that in the J-P treatment (two-sided Mann-Whitney test, p = .6471), but is significantly different from that in the I-P-P treatment (two-sided Mann-Whitney test, p = .0087).

What happens when there are two individual third-party punishers that act separately and independently toward the same targets? Each person in a third-party enforcement pair in the 2-I-P-S treatment imposed on average 4.00 points to a defector in Scenario (b), which is not significantly different from that in the I-P treatment. As discussed earlier, the only difference between the I-P-P and 2-I-P-S treatments is whether the targets of punishment are the same or not. Thus, the difference in the punishment per third-party individual between the I-P-P and 2-I-P-S treatments (6.63 points versus 4.00 points) can be interpreted as the free-riding behavior of third-party punishers in the 2-I-P-S treatment. Despite the free-riding behavior, the average punishment received by defectors in Scenario (b) is 8.00 points (= 4.00×2), which is even higher than that in the I-P-P treatment, because the number of punishers is two in the 2-I-P-S treatment.

RESULT 3: *The punishment intensity in Scenario (b) is much stronger in the I-P-P and 2-I-P-S treatments than in the I-P, J-P and J-P-C treatments.*

As a robustness check, we also perform a regression analysis by adding clustering to control for possible correlations within sessions (e.g., Fréchette 2012); which confirms RESULT 3. The dependent variable is the average punishment received per second-party player in Scenario (b). The independent variables include treatment dummy variables. As shown in Table 2, the estimation first confirms that defectors receive significantly stronger punishment in the I-P-P and 2-I-P-S treatments than in the I-P treatment. Second, the punishment intensity in the two joint-decision treatments is not significantly different from that in the I-P treatment.

4.3. How Deterrent is the Third-Party Punishment?

The seminar work of Fehr and Fischbacher (2004a) found that having a single third-party individual in a group cannot be a deterrent for a second-party player to select defection. A closer look at the data shows that making subjects' action choices known by third-party partners in the enforcement teams as in the I-P-P treatment or having two independent individual punishers as in the 2-I-P-S treatment raises the level of third-party punishment high enough that choosing to defect is no longer materially beneficial. At the same time, having only a single individual punisher in a group as in the I-P treatment (which is the same setup as Fehr and Fischbacher 2004a) or letting a pair of two individuals make a single punishment decision to a target as in the J-P and J-P-C treatments does not have the same effect.

We calculate two-party player *i*'s expected payoff when she selects to cooperate or defect in each treatment by following two steps (Table 3). First, we compute her payoff after the punishment stage for each of the four scenarios by using average realized third-party punishment points in Figure 1 (columns (1) to (4) in Table 3). Then, we compute her expected payoff using

the realized percentage of cooperators or defectors as weights of the payoffs in calculating the average payoff when she chooses to cooperate or defect by treatment (bold numbers in Table 3). We see that the expected payoffs when a subject chooses to cooperate are 13.7%, 12.1% and 9.5% lower than those when she chooses to defect in the I-P, J-P and J-P-C treatments, respectively. By contrast, the former rates are 19.4% and 23.8% *higher* than the latter rates in the I-P-P and 2-I-P-S treatments, respectively.

RESULT 4: *Individually decided third-party punishment is not strong enough to make defecting materially unbeneficial as is consistent with Fehr and Fischbacher (2004a) if there is only one individual third-party punisher in a group. Third-party punishment is, however, strong enough to do so if a subject's action choice is made known to her enforcement team partner in the I-P-P treatment. When punishment decisions are made jointly by two individuals in an enforcement team, the punishment intensity is too weak to prevent a second-party player from defecting.*

RESULT 5: *Third-party punishment is strong enough to prevent a two-party player from choosing to defect if there are two independent punishers in a group.*

In Section 4.4, we explore why the punishment intensity in the two joint-decision treatments is much lower than that in the I-P-P treatment.

4.4. The Impact of Communication in the J-P treatment

Each third-party punisher in the J-P treatment has a partner as in the I-P-P treatment. How does having a partner itself affect the intensity of third-party punishment in the J-P treatment? We can approximately measure the impact of having a partner in the J-P treatment by using the third-party players' willingness to punish which is elicited before the communication stage.[24]

[24] We acknowledge that the pre-communication willingness to punish in the J-P treatment is not perfectly comparable to the willingness to punish in the I-P-P treatment. As mentioned in Section 2, the pre-communication willingness to punish is elicited on condition that (1) subjects can unilaterally decide pair punishment points to a

Panel (i) of Figure 2 shows the average pre-communication willingness to punish in the J-P treatment along with the average actual punishment points after communication. We find two clear results. First, the average pre-communication willingness to punish under Scenario (b) in the J-P treatment is almost at the same level as the punishment imposed under that scenario in the I-P-P treatment (see Figures 1(ii) and 2(i)). The former (6.19 points) is not significantly different from the latter (6.63 points) according to a Mann-Whitney test (p = .5961, two-sided). This implies that the social effect of having a pair partner can be at work in the J-P treatment as in the I-P-P treatment. The same also holds for the companion treatment – the J-P-C treatment (see Figure 2(ii)). This suggests that the significance of such a social effect is robust to different social distance.

RESULT 6: *The pre-communication willingness to punish in Scenario (b) in the J-P and J-P-C treatments is not significantly different from punishment intensity imposed in Scenario (b) in the I-P-P treatment.*

Second, as shown in Figure 2, subjects' willingness to punish under Scenario (b) in the J-P and J-P-C treatments decreases substantially after communicating with pair partners. The decreases are statistically significant according to two-sided Wilcoxon signed ranks tests (p = .0171 for the J-P treatment; p = .0107 for the J-P-C treatment).[25] This suggests that a likely factor for the pairs' weak punishment despite RESULT 6 in the J-P and J-P-C treatments is the effects of communication within pairs. This is consistent with past research on team decision-

target and (2) the punishment decisions affect not only their own payoffs but also their pair partners' payoffs. The aspect (2) is absent in the I-P-P treatment as two third-party team members act independently and the payoffs depend only on their own punitive acts in the I-P-P treatment. Also, the targets of two persons in a team are different in the I-P-P treatment, while they are the same in the J-P treatment. Differently speaking, if the level of the pre-communication willingness to punish in the J-P treatment is different from the willingness to punish in the I-P-P treatment, it may be attributable to the effect of aspect (2) or the differences in the structure of punishment targets.

[25] See Appendix Table A.7.

making that has suggested that letting a pair jointly make a single action choice strengthens their strategic behavior as discussed in Section 3.

Does the impact of communication differ by pairs' punitive disposition? Lastly, we take a look at the data by classifying a pair "self-regarding" if the pair's average pre-communication willingness to punish is below the session average and classify one "other-regarding" if the average pre-communication willingness to punish is above the session average as is similar to Cason and Mui (1992) (and also Kamei (2016)). As shown in Table 4, nine (eight) pairs are classified as self-regarding and eight (nine) pairs are classified as other-regarding in the J-P (J-P-C) treatment. The data indicates that the other-regarding pairs' willingness to punish under Scenario (b) decreases substantially after the communication in both the J-P and J-P-C treatments. In addition, we observe a significant decrease in self-regarding pairs' punishment intensity through communication in the J-P treatment (the punishment intensity was sufficiently small for the self-regarding pairs even before communication in the J-P-C treatment). This finding is similar to Kamei (2016) that shows that although two-person pairs become more forward-looking than individuals, the pairs do not sustain cooperation with other peers in a dilemma, as is the case for individuals, when they do not have opportunities to build reputations in a random-matching environment.

RESULT 7: *Other-regarding pairs decrease their willingness to punish through the joint decision-making process significantly in both the J-P and J-P-C treatments. Self-regarding pairs also do so in the J-P treatment.*

5. Conclusions

This paper experimentally explored how the strength of third-party punishment differs by the third party's decision-making format. Our data first replicates the well-known finding in the

literature that third-party punishment is not a deterrent to second-party players if there is only one individual third-party player in a group. The non-deterrent strength of punishment is also observed when two third-party punishers are faced with the same targets and jointly decide single punishment points to them. The data, however, shows that third-party punishment acts as a sufficient deterrent against second parties' defection when each third-party punisher is paired with another punisher in a different group and her punishment act is shared with the paired partner, or when there are two independent individual third-party punishers in the same group with two-party players.

Lastly, we note that there are many future research possibilities, considering the third parties' sensitivity to decision-making formats we found. For example, we found free-riding behaviors between two individual punishers in the 2-I-P-S treatment although the third-party punishment is deterrent in that treatment. How deterrent would third-party punishment be if there are more than two independent punishers in a group? Past research on collective action dilemmas (e.g., public goods dilemmas) suggests that people's free-riding behaviors could depend on the group size (e.g., Isaac and Walker 1988). Considering that third parties' free-riding behavior could also depend on the number of third parties in a group, the third-party punishment intensity may or may not be sufficiently a deterrent when more than two punishers are present. For another example, a number of decision-making formats co-exist in reality. It would also be interesting to explore how cooperation norms could evolve in a society given the distribution of different formats in the real world. Also, our study excludes a possibility that people select the format of third-party enforcement themselves; and how they endogenously implement a decision-making format could be another interesting area for further research. People's endogenous selection of third-party punishment formats and its efficiency could also differ by

context and purpose (e.g., everyday life, legal enforcement such as courts, police enforcement, vigilante).

REFERENCES

Ariely, D., Bracha, A., Meier, S., 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* 99: 544-555.

Auerswald, H., Schmidt, C., Thum, M., Torsvik, G., 2016. "Teams Contribute More and Punish Less." working paper.

Bernhard, H., Fehr, E., Fischbacher, U., 2006. " Group Affiliation and Altruistic Norm Enforcement." *American Economic Review* 96: 217-221.

Bock, O., Nicklisch, A., Baetge, I., 2014. "hroot: Hamburg Registration and Organization Online Tool." *European Economic Review* 71: 117-120.

Bowles, S., Gintis, H., 2005. "Prosocial emotions," in L. Blume, S. Durlauf (Eds.), The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow, Oxford University Press, Oxford: 337-367

Buckholtz, J., Asplund, C., Dux, P., Zald, D., Gore, J., Jones, O., Marois, R., 2008. "The Neural Correlates of Third-Party Punishment." *Neuron* 60: 930-940.

Carpenter, J., Matthews, P., 2012. "Norm Enforcement: Anger, Indignation, or Reciprocity?" *Journal of European Economic Association* 10: 555-572.

Cason, T., Mui, V.-L., 1997. "A Laboratory Study of Group Polarisation in the Team Dictator Game." *Economic Journal* 107: 1465-1483.

Charness, G., Sutter, M., 2012. "Groups Make Better Self-Interested Decisions." *Journal of Economic Perspective* 26: 157-176.

Chaudhuri, A., 2011. "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature." *Experimental Economics* 14:47-83.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R., Fink, G., 2013. "Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study." *Social Cognitive and Affective Neuroscience* 8: 424-431.

Ertan, A., Page, T., Putterman, L., 2009. "Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free-Rider Problem." *European Economic Review* 53: 495-511.

Fehr, E., Fischbacher, U., 2004a. "Third-party Punishment and Social Norms." *Evolution and Human Behavior* 25: 63-87.

Fehr, E., Fischbacher, U., 2004b. "Social norms and human cooperation." *Trends in Cognitive Sciences* 8: 185-190.

Fehr, E., Gächter, S., 2002. "Altruistic punishment in humans." *Nature* 415: 137-140.

Fehr, E., Schmidt, K., 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114: 817-68.

Fehr, E., Schmidt, K., 2006. ''The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories.'' In *Handbook of the Economics of Giving, Altruism and Reciprocity*, edited by S.-G. Kolm and J. M. Ythier, pp. 615-91. North Holland.

Feri, F., Irlenbusch B., Sutter, M., 2010. "Efficiency Gains from Team-Based Coordination—Large-Scale Experimental Evidence." *American Economic Review* 100:1892-1912.

Fischbacher, U., 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10: 171-178.

Fréchette, G., 2012. "Session-effects in the laboratory." *Experimental Economics* 5:485-498.

Gächter, S., Renner, E., 2010. "The effects of (incentivized) belief elicitation in public goods experiments." *Experimental Economics* 13: 364-377.

Gillet, J., Schram, A., Sonnemans, J., 2009. "The Tragedy of the Commons Revisited: The Importance of Group Decision-Making." *Journal of Public Economics* 93: 785-797.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., 2006. "Costly punishment across human societies." *Science* 1767-1770.

Isaac, M., Walker, J., 1988. "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism." *Quarterly Journal of Economics* 103: 179-199.

Kamei, K., 2014. "Conditional Punishment." *Economics Letters* 124: 199-202.

Kamei, K., 2016. "Joint Decision-Making and Strategic Reputation Building in a Finitely-Repeated Dilemma." working paper.

Kamei, K., Putterman, L., 2015. "In broad daylight: Fuller information and higher-order punishment opportunities can promote cooperation." *Journal of Economic Behavior and Organization* 120: 145-159.

Kamei, K., Putterman, L., Tyran, J.-R. 2015. "State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods." *Experimental Economics* 18: 38-65.

Kerr, N., Tindale, R.S., 2004. "Group Performance and Decision Making." *Annual Review of Psychology* 55: 623-655.

Kocher, M., Sutter, M., 2007. "Individual versus group behavior and the role of the decision making procedure in gift-exchange experiments." *Empirica* 34: 63-88.

Kugler, T., Kausel, E., Kocher, M., 2012. "Are groups more rational than individuals? A review of interactive decision making in groups." *WIREs Cognitive Science* 3: 471-482.

Kurzban, R., DeScioli, P., O'Brien, E., 2007. "Audience effects on moralistic punishment." *Evolution and Human Behavior* 28: 75-84.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., Sutter, M., 2014. "Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation." *Proceedings of the National Academy of Sciences* 111: 6916-6921.

Ledyard, J., 1995. "Public goods: A survey of experimental research." In J. H. Kagel and A.E. Roth (eds.), The Handbook of Experimental Economics 111-194, Princeton University Press.

Leibbrandt, A., López-Pérez, R., 2012. "An exploration of third and second party punishment in ten simple games." *Journal of Economic Behavior and Organization* 84: 753-766.

Levine, D., 1998. "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics* 1: 593-622.

Lewisch, P., Ottone, S., Ponzano, F., 2011. "Free-Riding on Altruistic Punishment? An Experimental Comparison of Third-Party Punishment in a Stand-Alone and in an In-Group Environment." *Review of Law and Economics* 7: 165-194.

Lieberman, D., Linke, L., 2007. "The effect of social category on third party punishment." *Evolutionary Psychology* 5: 289-305.

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.

Marcin, I., Robalo, P., Tausch, F., 2016. "Institutional Endogeneity and Third-party Punishment in Social Dilemmas." Max Planck Institute for Research on Collective Goods working paper.

Marlowe, F., Berbesque, C., Barrett, C., Bolyanatz, A., Gurven, M., Tracer, D., 2010. "The 'spiteful' origins of human cooperation." *Proceedings of the Royal Society B* 278: 2159-2164.

McAuliffe, K., Jordan, J., Warneken, F., 2015. "Costly third-party punishment in young children." *Cognition* 134: 1-10.

Müller, W., Tan, F., 2013. "Who acts more like a game theorist? Group and individual play in a sequential market game and the effect of the time horizon." *Games and Economic Behavior* 82: 658-674.

Nelissen, R., Zeelenberg, M., 2009. "Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions." *Judgment and Decision Making* 4: 543-553.

Putterman, L., Tyran, J.-R., Kamei, K., 2011. "Public goods and voting on formal sanction schemes." *Journal of Public Economics* 95: 1213-1222.

Riedl, K., Jensena, K., Calla, J., Tomaselloa, M., 2012, "No third-party punishment in chimpanzees." *Proceedings of the National Academy of Sciences* 109: 14824-14829.

Samek, A., Sheremeta, R., 2014. "Recognizing contributors: an experiment on public goods." *Experimental Economics* 7: 673-690.

Sell, J., Wilson, R., 1991. "Levels of Information and Contributions to Public Goods." Social Forces 70: 107-124.

Shang, J., Croson, R., 2009. "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods." *Economic Journal* 119: 1422-1439

Sobel, J., 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature* 43: 392-436.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, R., Lis, S., Windmann, S., Kirsch, P., 2011. "Beyond revenge: Neural and genetic bases of altruistic punishment." *NeuroImage* 54*:* 671-680.

Sutter, M., Lindner, P., Platsch, D., 2009. "Social norms, third-party observation and third-party reward." *University of Innsbruck Working Papers in Economics and Statistics*, No. 2009-08.

van Miltenburg, N., Buskens, B., Vincent, D., Raub, W., 2014. "Implementing punishment and reward in the public goods game - The effect of individual and collective decision rules." *International Journal of the Commons* 8: 47-78.

Van Lange, P., Rockenbach, B., Yamagishi. T. (editors), 2011. *Reward and Punishment in Social Dilemmas*. Oxford University Press.

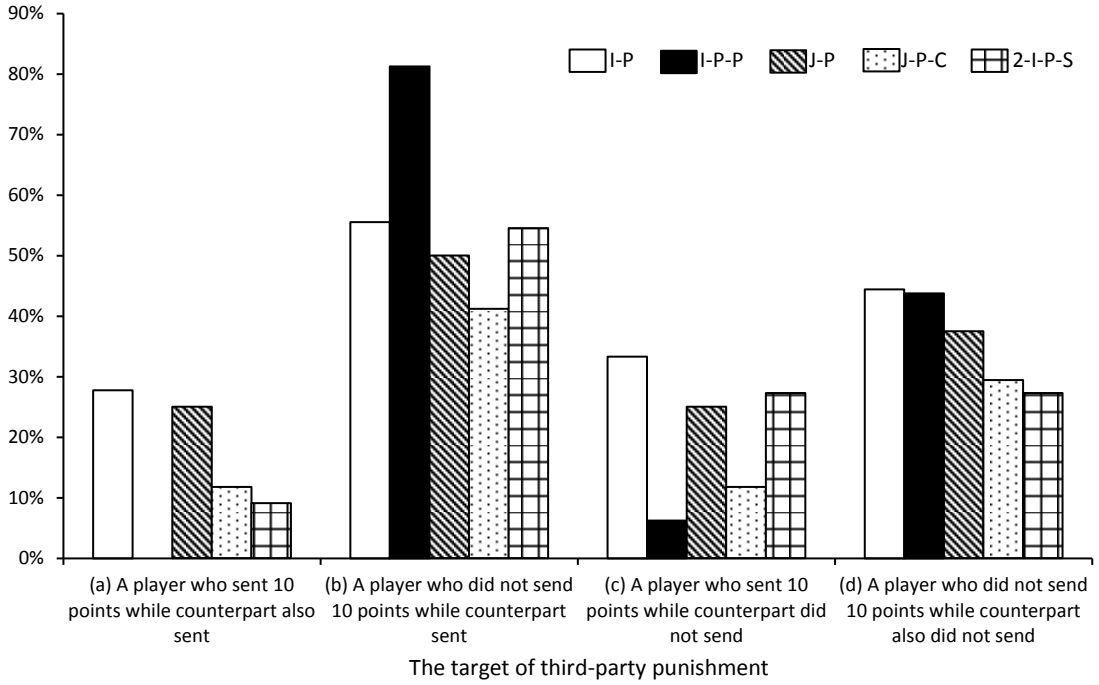
Weimann, J., 1994. "Individual Behavior in a Free Riding Experiment." *Journal of Public Economics* 54: 185-200.

Wilson, R., Sell, J., 1997. ""Liar, Liar …" Cheap Talk and Reputation in Repeated Public Goods Settings." *Journal of Conflict Resolution* 41: 695-717.

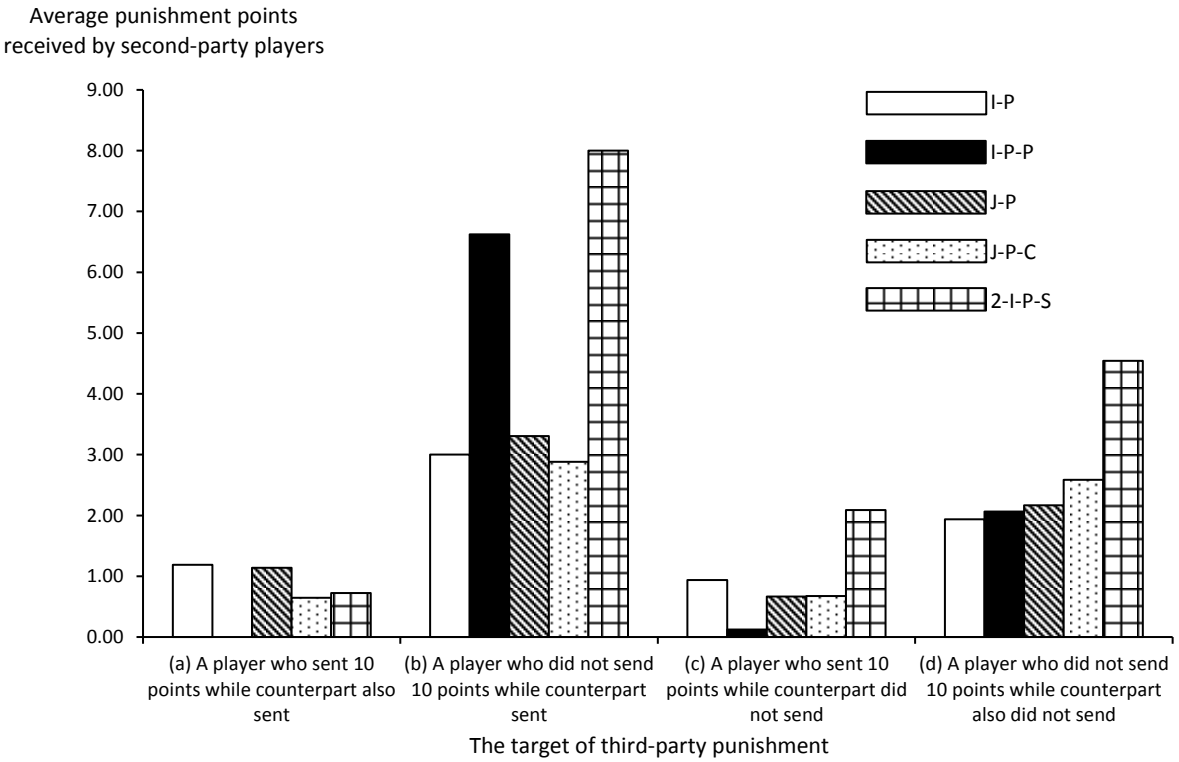
Table 1: *Summary of Treatments*

Treatment name	Third party	Subjects' punishment decisions	Decision-making protocol (individual or joint decision)	Number of subjects
I-P (Individual Punishment)	Individual	---	Individual	48 subjects
I-P-P (Individual Punishment with Partner)	Pair	Informed to the pair partner	Individual (two third-party individuals independently impose punishment in different groups.)	48 subjects
J-P (Joint Punishment)	Pair	Informed to the pair partner	Joint (each third-party player is not aware of who their partner is in their pair.)	72 subjects
2-I-P-S (2 Individual Punishment towards the Same Target)	Pair	Informed to the pair partner	Individual (two third-party individuals in a group simultaneously and independently impose punishment to the same second-party players.)	44 subjects
[Robustness check of the J-P treatment]				
J-P-C (Joint Punishment, Close Social Distance)	Pair	Informed to the pair partner	Joint (each third-party player is aware of who their partner is in their pair.)	68 subjects
Total	---	---	---	280 subjects

Figure 1: *Frequency of Third-Party Punishment, and Average Third-Party Punishment Points Received, by Treatment and Scenario*



(i) Frequency of Third-Party Punishment



(ii) Average Third-Party Punishment Points Received

Table 2: *Third-party Punishment Received under Scenario (b) by Treatment*

Dependent variable: Total punishment points received by a second-party defector *i* in Scenario (b)

Estimation Method:	Tobit regression	Ordered Probit regression
Independent Variable:	(1)	(2)
(i) The I-P-P dummy {= 1 for the I-P-P treatment; and 0 otherwise}	5.19*** (1.17)	.83*** (.25)
(ii) The J-P dummy {= 1 for the J-P treatment; and 0 otherwise}	-.87 (2.11)	-.056 (.30)
(iii) The J-P-C dummy {= 1 for the J-P-C treatment; and 0 otherwise}	-1.49 (1.77)	-.19 (.23)
(iv) The 2-I-P-S dummy {= 1 for the 2-I-P-S treatment; and 0 otherwise}	7.02** (3.06)	.97** (.42)
Constant term	.92 (1.09)	----[#1]
# of observations	70	70
Log Pseudo likelihood	-159.7	-150.7

Notes: Standard errors, clustered by session, are in parentheses. The reference group is the punishment points received in Scenario (b) in the I-P treatment. The number of left-(right-) censored observations is 28(2) in column (1). See Appendix Table A.5 for F (Chi-squared) test results for comparisons between the coefficient estimates in column (1) (column (2)).

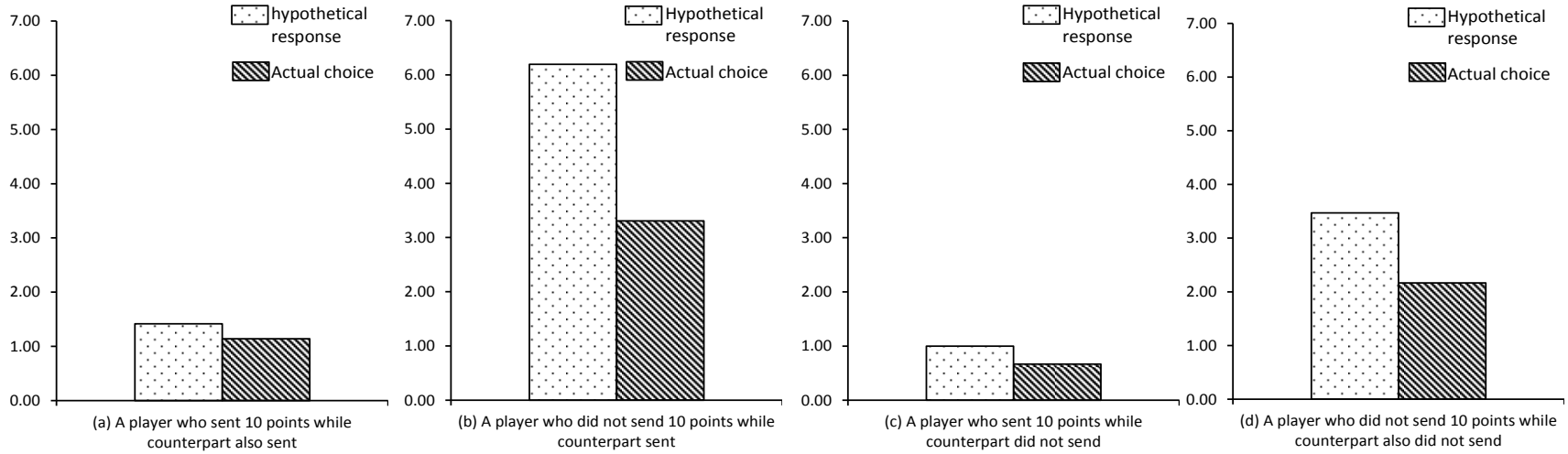
[#1] The estimates of cut points are omitted to conserve space.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

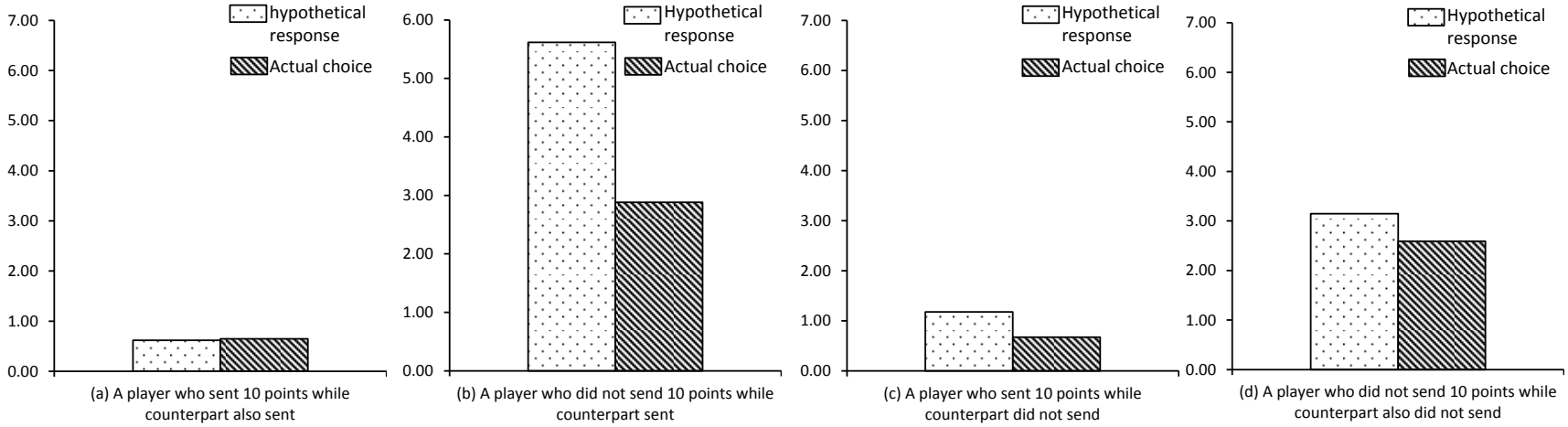
Table 3: *Average Expected Payoff of Subject i when Choosing to Send or Not to Send 10 points*

Subject *i*'s action choice:	Send (Cooperate)			Not to send (Defect)		
Subject *i*'s counterpart *j*'s action choice:		Send Scenario (a) (1)	Not to send Scenario (c) (2)		Send Scenario (b) (3)	Not to send Scenario (d) (4)
(i) Subject *i*'s payoff in Stage 1	----	45	15	----	55	25
(ii) Subject *i*'s payoff in Stage 1 (line (i)) minus average punishment amount subject *i* would receive (see Figure 1(ii))						
I-P treatment	----	41.44	12.19	----	46.00	19.19
I-P-P treatment	----	45.00	14.63	----	35.13	18.81
J-P treatment	----	41.58	13.00	----	45.08	18.50
J-P-C treatment	----	43.06	12.97	----	46.35	17.24
2-I-P-S treatment	----	42.82	8.73	----	31.00	11.36
(iii) Subject *i*'s expected payoff in Stage 2 [line (ii) × the % of cooperators or defectors for columns (1) to (4)]						
I-P treatment	**33.21**	29.78	3.43	**38.46**	33.06	5.40
I-P-P treatment	**36.46**	32.34	4.11	**30.54**	25.25	5.29
J-P treatment	**30.86**	25.99	4.88	**35.11**	28.18	6.94
J-P-C treatment	**34.21**	30.39	3.81	**37.79**	32.72	5.07
2-I-P-S treatment	**27.32**	23.36	3.97	**22.07**	16.91	5.17

Figure 2: *Pairs' Pre-communication Willingness to Punish versus After-Communication Willingness to Punish*



(i) The J-P treatment



(ii) The J-P-C treatment

Notes: The vertical axes indicate average punishment points per two-party player. Each figure shows pre-communication and after-communication punishment points. See Appendix Table A.6 for comparisons between the hypothetical responses and actual punishment in the I-P or I-P-P treatment.

Table 4: *Shifts in the Third-Party Pairs' Willingness to Punish through Communication*

	J-P treatment		J-P-C treatment	
	Self-regarding pair (1)	Other-regarding pair (2)	Self-regarding pair (3)	Other-regarding pair (4)
The number of pairs	9[#1]	8[#1]	8[#2]	9[#2]
(i) Average pre-communication punishment points in Scenario (b)	3.61	9.44	2.25	8.61
(ii) Average after-communication punishment points in Scenario (b)	1.39	5.63	3.13	2.67
The Direction of Shifts				
Upward shift ($C_i > C_i^{pre}$)	2	2	2	1
No shift ($C_i = C_i^{pre}$)	1	0	2	0
Downward shift ($C_i < C_i^{pre}$)	6	6	4	8
Wilcoxon signed ranks test for the null H_0: (i) = (ii)				
p-value (two-sided)	.0363**	.0793*	.6180	.0105**

Notes: [#1] The total number of pairs in the J-P treatment is 18; and one pair's pre-communication average punishment points in Scenario (b) is the same as the session average – so this pair was not assigned either self-regarding or other-regarding pair. [#2] The total number of pairs in the J-P-C treatment is 17.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

Appendix A: Additional Figures and Tables

Table A.1: *Two-Party Players' Sending Rates and Third-party Punishers' Beliefs on the Number of Cooperators (supplementing Section 4.1 of the paper)*

(1) Test Results for the Differences in Sending Rate of Second-party Players between Treatments

	I-P	I-P-P	J-P	J-P-C	2-I-P-S
I-P	---	1.000	.2107	1.000	.2498
I-P-P	---	---	.2107	1.000	.2498
J-P	---	---	---	.2235	1.000
J-P-C	---	---	---	---	.2620
2-I-P-S	---	---	---	---	---

Notes: Two-sided Fisher's exact tests. The numbers in this table are *p*-values.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

(2) Test Results for the Differences in Third-party Punishers' Beliefs between Treatments

	I-P	I-P-P	J-P	J-P-C	2-I-P-S
I-P	---	.6308	.8878	.7059	.9232
I-P-P	---	---	.6858	.9173	.5592
J-P	---	---	---	.7566	.7932
J-P-C	---	---	---	---	.6192
2-I-P-S	---	---	---	---	---

Notes: Two-sided Mann-Whitney tests. The numbers in this table are *p*-values.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

Table A.2: *The Differences in the Frequency of Third-Party Punishment between Scenarios (supplementing Figure 1(i) of the paper)*

(I) All treatments (tests using all data)

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0001***	.3446	0.0122**
Scenario (b)	---	---	.0016***	.1096
Scenario (c)	---	---	---	.1148
Scenario (d)	---	---	---	

(II) By treatment

a. The I-P Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.3603	1.0000	.5322
Scenario (b)	---	---	.5487	.7773
Scenario (c)	---	---	---	.7572
Scenario (d)	---	---	---	

b. The I-I-P Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0014***	1.0000	.0287**
Scenario (b)	---	---	.0073***	.3918
Scenario (c)	---	---	---	.1074
Scenario (d)	---	---	---	---

c. The J-P Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.4982	1.0000	.7225
Scenario (b)	---	---	.4982	.7539
Scenario (c)	---	---	---	.7225
Scenario (d)	---	---	---	---

d. The J-P-C Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.2575	1.0000	.4192
Scenario (b)	---	---	.2575	.7419
Scenario (c)	---	---	---	.4192
Scenario (d)	---	---	---	---

e. The 2-I-P-S Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0279**	.2615	.2615
Scenario (b)	---	---	.2716	.2716
Scenario (c)	---	---	---	1.000
Scenario (d)	---	---	---	---

Notes: Two-sided Fisher's exact tests. The numbers in these tables are two-sided *p*-values.
*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

Table A.3: *Non-Parametric Tests for the Differences in Average Punishment Point between Scenarios (supplementing Figure 1(ii) of the paper)*

(I) The I-P Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0394**	.6045	.1716
Scenario (b)	---	---	.0367**	.4188
Scenario (c)	---	---	---	.0481**
Scenario (d)	---	---	---	

(II) The I-I-P Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0007***	.3173	.0088***
Scenario (b)	---	---	.0009***	.0021***
Scenario (c)	---	---	---	.0149**
Scenario (d)	---	---	---	---

(III) The J-P Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0084***	.1573	.0608*
Scenario (b)	---	---	.0006***	.0158**
Scenario (c)	---	---	---	.0048***
Scenario (d)	---	---	---	---

(IV) The J-P-C Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0010***	.9832	.0530*
Scenario (b)	---	---	.0016***	.1874
Scenario (c)	---	---	---	.0530*
Scenario (d)	---	---	---	---

(V) The 2-I-P-S Treatment

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
Scenario (a)	---	.0007***	.0993*	.0501*
Scenario (b)	---	---	.0050***	.0119**
Scenario (c)	---	---	---	.3829
Scenario (d)	---	---	---	---

Notes: Two-sided Wilcoxon signed ranks tests. The numbers in these tables are two-sided *p*-values.
*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

Table A.4: *Non-Parametric Tests for the Differences in the Average Punishment Point Received by Second-Party Players between Treatments (supplementing Figure 1(ii) of the paper)*

(1) Testing results for the equality of punishment received by cooperators in Scenario (a)

	I-P	I-P-P	J-P	J-P-C	2-I-P-S
I-P	---	.0841*	.5933	.1899	.4603
I-P-P	---	---	.1570	.3221	.2153
J-P	---	---	---	.4069	.7995
J-P-C	---	---	---	---	.6429
2-I-P-S	---	---	---	---	---

Notes: Two-sided Mann-Whitney tests. The numbers in this table are *p*-values.
*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

(2) Test results for the equality of punishment received by defectors in Scenario (b)

	I-P	I-P-P	J-P	J-P-C	2-I-P-S
I-P	---	.0182**	.7982	.5271	.0261**
I-P-P	---	---	.0725*	0.0087***	1.0000
J-P	---	---	---	.6471	0.0264**
J-P-C	---	---	---	---	.0162**
2-I-P-S	---	---	---	---	---

Notes: Two-sided Mann-Whitney tests. The numbers in this table are *p*-values.
*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

(3) Test results for the equality of punishment received by cooperators in Scenario (c)

	I-P	I-P-P	J-P	J-P-C	2-I-P-S
I-P	---	.1693	.3912	.1138	.2111
I-P-P	---	---	.4673	.9589	.0448**
J-P	---	---	---	.4213	.0496**
J-P-C	---	---	---	---	.0214**
2-I-P-S	---	---	---	---	---

Notes: Two-sided Mann-Whitney tests. The numbers in this table are *p*-values.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

(4) Test results for the equality of punishment received by defectors in Scenario (d)

	I-P	I-P-P	J-P	J-P-C	2-I-P-S
I-P	---	.4446	.6987	.4136	.4470
I-P-P	---	---	.3636	.1395	.8319
J-P	---	---	---	.8726	.2682
J-P-C	---	---	---	---	.1993
2-I-P-S	---	---	---	---	---

Notes: Two-sided Mann-Whitney tests. The numbers in this table are *p*-values.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

RESULT: (1) Defectors in Scenario (b) receive significantly (weakly significantly) stronger punishment in the I-P-P treatment than in the I-P and J-P-C treatment (in the J-P treatment).

(2) Defectors in Scenario (b) receive significantly stronger punishment in the 2-I-P-S treatment than in the I-P, J-P and J-P-C treatments.

(3) The average punishment points received in Scenario (b) are not significantly different between the I-P-P treatment and the 2-I-P-S treatment.

Table A.5: *Comparing the Coefficient Estimates of Table 2 of the Paper*

(a) F tests for the equality of coefficient estimates in column (1)

	Variable (i)	Variable (ii)	Variable (iii)	Variable (iv)
Variable (i)	---	.0015***	.0000***	.5252
Variable (ii)	---	---	.7763	.0243**
Variable (iii)	---	---	---	.0081***
Variable (iv)	---	---	---	---

Notes: The numbers are *p*-values.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

(b) Wald chi-squared tests for the equality of coefficient estimates in column (2)

	Variable (i)	Variable (ii)	Variable (iii)	Variable (iv)
Variable (i)	---	.0015***	.0000***	.7208
Variable (ii)	---	---	.6386	.0172**
Variable (iii)	---	---	---	.0053***
Variable (iv)	---	---	---	---

Notes: The numbers are *p*-values.

*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

Table A.6: *Pre-Communication Willingness to Punish in the J-P and J-P-C treatments versus Punishment Intensity in the I-P and I-P-P treatments in Scenario (b) (supplementing Figure 2 of the paper)*

(1) Test Results for the Null Hypothesis: Pre-communication Punishment Points in the Joint-Decision Treatment = Actual Punishment Points in the I-P treatment

	J-P treatment versus I-P treatment	J-P-C treatment versus I-P treatment
p-value (two-sided)	.0303**	.0562*

Notes: Mann-Whitney test results. *, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

(2) Test Results for the Null Hypothesis: Pre-communication Punishment Points in the Joint-Decision Treatment = Actual Punishment Points in the I-P-P treatment

	J-P treatment versus I-P-P treatment	J-P-C treatment versus I-P-P treatment
p-value (two-sided)	.5961	.3058

Notes: Mann-Whitney test results. *, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

RESULT: *Individuals in two-person pairs in the J-P (J-P-C) treatment exhibit significantly (weakly significantly) stronger willingness to punish defectors in Scenario (b) before communication, compared with individuals in the I-P treatment. The levels of pre-communication willingness to punish in the J-P and J-P-C treatment are at similar levels to the I-P-P treatment.*

Table A.7: *Decreases in Willingness to Punish through Communication in the J-P and J-P-C treatments (supplementing Figure 2 of the paper)*

	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)
J-P treatment				
Actual punishment points *minus* Pre-communication punishment points	–0.28 points	–2.89 points	–0.33 points	–1.31 points
p-value (two-sided)[#1]	.9515	.0171**	.1888	.0995*
J-P-C treatment				
Actual punishment points *minus* Pre-communication punishment points	0.03 points	–2.74 points	–0.50 points	–0.56 points
p-value (two-sided) [#1]	1.0000	.0107**	.6350	.0873*

Notes: [#1] *p*-values for Wilcoxon signed ranks tests. *, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

RESULT: *Subjects in the two joint-decision treatments decrease willingness to punish in Scenario (b) significantly through communication. They also decrease willingness to punish in Scenario (d) weakly significantly through communication.*