"Benevolent"or "Wise": The Effect of Third-Party Punishment on Punisher Reputation

Authors: Sijing Chen, Yechao Xu, Chen Sijing

Date: 2020-09-12T00:00:00+00:00

Abstract

Third-party punishment exerts significant influence on the punisher' s reputation; however, existing literature offers divergent answers regarding the direction of this effect. A potential reason underlying this issue is that prior research has failed to differentiate between distinct dimensions of reputation and the varied motivations and forms of punishment. By incorporating the warmth-competence two-dimensional framework into the punisher' s reputation, experimental results demonstrate that third-party punishment generally diminishes evaluations of the punisher on the warmth dimension while enhancing evaluations on the competence dimension. Moderation analysis reveals that punishment with motivation attributed to collective focus further amplifies its positive impact on competence and attenuates its negative impact on warmth; moreover, the higher the punisher' s level of cooperation, the greater the extent to which their motivation is attributed to collective focus. Further analysis targeting different punishment forms indicates that when punishment motivation is attributed to individual focus, economic punishment exerts a significantly more negative effect on warmth than social punishment, whereas under collective focus attribution, economic punishment produces a significantly less positive effect on competence than social punishment.

Full Text

Preamble

"The Benevolent" or "The Wise" : The Impact of Third-Party Punishment on Punishers' Reputation

Chen Sijing, Xu Yechao
(School of Economics and Management, Zhejiang University of Science and Technology, Hangzhou, 310023, China)

Abstract

Third-party punishment significantly influences the punisher' s reputation, yet existing literature offers divergent answers regarding the direction of this effect. A potential reason for this discrepancy is that previous studies have failed to distinguish between different dimensions of reputation and varied motivations and forms of punishment. By introducing the warmth-competence dual-dimensional framework to the punisher' s reputation, experimental results demonstrate that third-party punishment generally reduces evaluations of the punisher on the warmth dimension while enhancing evaluations on the competence dimension. Moderation analysis reveals that punishment attributed to collective-focused motives further amplifies its positive effect on competence and mitigates its negative effect on warmth. Moreover, the higher the punisher' s cooperation level, the more their motives are attributed to collective focus. Further analysis of different punishment forms shows that when punishment motives are attributed to individual focus, financial sanctions have a significantly stronger negative effect on warmth than social sanctions, whereas under collective-focused attribution, financial sanctions have a significantly weaker positive effect on competence than social sanctions.

Keywords: third-party punishment; social norm; punishment motive; reputation; financial sanction; social sanction

Classification: B849: C91

1. Introduction

In social science literature, cooperation is typically defined as behavior where individuals incur costs to benefit others (Rand, 2016). Cooperation is crucial for solving numerous problems in human society (Bear & Rand, 2016), and we have developed norms to facilitate it (de Kwaadsteniet et al., 2007; Fehr & Schurtenberger, 2018). However, compliance with cooperative norms is not automatic, as individuals tend to pursue self-interest, which often leads to insufficient provision of public goods and losses in social efficiency (de Kwaadsteniet et al., 2019). How, then, is large-scale cooperation among non-kin individuals sustained? Fehr and Gächter' s (2002) third-party punishment theory offers a partial explanation, suggesting that certain individuals have an innate tendency to punish norm violators. As long as a sufficient number of such individuals exist, cooperative relationships among group members can be maintained (Carpenter et al., 2009). Yet because the costs of third-party punishment (money, time, energy, and potential retaliation) are borne by the punisher while the benefits are shared by all group members, third-party punishment creates a second-order social dilemma (Colman, 2006; Hauert et al., 2007): compared to punitive cooperators (hereafter "punishers"), individuals who cooperate but do not punish are second-order free-riders. Since punishment costs are shouldered by punishers, second-order free-riders necessarily have higher evolutionary fitness than punishers (Xie et al., 2017; Hu et al., 2016), raising a new question:

how do punishers prevail in evolution?

One widely discussed perspective is that third-party punishment brings positive reputation to punishers (Barclay, 2006; Barclay & Kiyonari, 2014), which in turn yields corresponding benefits, such as increased probability of receiving help from others in future interactions (Santos et al., 2010) or signaling that the punisher possesses desirable qualities (Jordan et al., 2016). If these benefits exceed punishment costs, punishers can be selected for in evolution. This view, primarily based on indirect reciprocity theory or costly signaling theory, provides a theoretical explanation for the evolution of third-party punishment and has received some empirical support (e.g., Jordan & Rand, 2019; Kurzban et al., 2007). However, its premise is that the punisher' s reputation must be positive. Yet growing evidence suggests that punishers' reputation is not necessarily positive (Bornstein & Weisel, 2010) and may even be negative (de Kwaadsteniet et al., 2019; Ozono & Watabe, 2012). Moreover, punishment does not necessarily increase the probability of receiving help from others (Kiyonari & Barclay, 2008). This indicates that the relationship between third-party punishment and punisher reputation may be more complex than anticipated, necessitating a deeper examination of the reputation mechanism to effectively explore whether reputation can fully explain the evolutionary advantages of third-party punishment.

Rand and Nowak (2013) noted that mainstream evolutionary theories of cooperation tend to simplify individuals as agents without motives, completely ignoring the importance of psychological motivation. This may be because current literature on cooperation and punishment primarily originates from economics, biology, and game theory (Chen & Yang, in press), while the absence of a psychological perspective has led us to largely overlook the role of motivation in the punisher reputation mechanism. In fact, people always make moral judgments about others' behavior based on motives, which subsequently affects interpersonal interactions (Bigman & Tamir, 2016). This means the same behavior can have vastly different effects on relationships depending on motivational attribution. Regarding punishment, studies by Fehr and Rockenbach (2003) and Liu and Xin (2014) have confirmed that motivational attribution significantly affects punishment' s impact on cooperation: only when punishment motives are attributed as altruistic can third-party punishment promote cooperation among the punished; otherwise, punishment inhibits cooperative behavior. A reasonable speculation is that punishment motives have a similar mechanism for punisher reputation—that is, only third-party punishment with reasonable motives can enhance the punisher's reputation. Based on this reasoning, we propose Research Question 1: Does attribution of third-party punishment motives significantly affect the punisher' s reputation?

Second, previous researchers have tended to treat reputation as a unidimensional variable, ignoring its different dimensions, which has resulted in punisher reputation being either entirely positive or entirely negative in prior studies (de Kwaadsteniet et al., 2019). As Beersma and van Kleef (2011) point out, reputation is essentially individuals' perception or evaluation of

others, and an important finding in relevant literature is that people typically use two fundamental dimensions to form evaluations of others (Fiske et al., 2007): warmth and competence. Warmth refers to benevolent traits exhibited in interactions with others, such as trustworthiness, while competence refers to one' s ability to achieve intended goals, such as action efficiency. In Chinese classical literature, Confucius' s question "What is a wise person like, what is a benevolent person like?" (*Xunzi · Zidao*) partially reflects this distinction; Wei Zheng in *Ten Reflections for Emperor Taizong* also addressed this point: "The wise exert their strategies···the benevolent spread their kindness." In real life, individuals highly evaluated on the warmth dimension may not receive the same evaluation on the competence dimension—the "nice guy" is a typical example, and vice versa. We believe Fiske et al.' s (2007) dual-dimensional theory of reputation also applies to third-party punishment, leading to Research Question 2: Does third-party punishment have different effects on the two dimensions of punisher reputation? In other words, does third-party punishment simultaneously affect both dimensions of reputation? Are the direction and magnitude of these effects consistent? Answering these questions will allow us to more finely reveal the different pathways through which punishment affects reputation.

Finally, third-party punishment in laboratory settings predominantly employs financial sanctions (Chen et al., 2014), where punishers pay monetary costs to reduce violators' earnings (Balliet et al., 2011). Although the payment of monetary costs often takes different forms across studies (Chen et al., 2020), Guala (2012) points out that this form of punishment is likely an artificial construct of laboratory settings. In real life, people prefer to use social sanctions to maintain norm enforcement. Social sanctions, also termed moral punishment (Cui et al., 2017), non-monetary punishment (Noussair & Tucker, 2005), or gossip (Wu, Balliet et al., 2016), essentially involve expressing moral condemnation of norm violations through verbal means without incurring monetary or material costs (Nelissen & Mulder, 2013; Noussair & Tucker, 2005). While some scholars have begun examining the effects of financial and social punishment on cooperation or social norms, no study has tested the impact of punishment form on punisher reputation, and existing literature on punisher reputation is largely based on financial punishment, potentially yielding one-sided conclusions. Since financial and social punishment differ significantly in manifestation (material deduction vs. verbal condemnation), cost (material vs. non-material), impact on individual outcomes (reducing violators' material benefits vs. reducing their reputation in the group), and impact on group outcomes (reducing group net benefits vs. not affecting group net benefits) (Guala, 2012), we hypothesize these two punishment forms also differentially affect punisher reputation. Thus, Research Question 3 is: Do financial and social punishment have different effects on punisher reputation?

2.1 Participants

We used G*Power 3.1 to determine the required sample size: with a medium effect size of $f^2 = 0.15$ and significance level $\alpha = 0.05$, 89 participants were needed to achieve 95% statistical power (1-β). The actual participants in Experiment 1 were 90 undergraduate students from a university who were not psychology majors. Participants had a mean age of 20.86 $\pm$ 1.27 years, with 61.11% being female. All participants had never participated in similar experiments. The distribution of majors was as follows: 36.67% science and engineering, 33.33% social sciences, 22.22% humanities, and 7.78% arts and other fields. Before the experiment began, we ensured participants fully understood the experimental rules and accurate meanings of technical terms through instructions and practice questions (examples in the appendix, same below), and obtained informed consent from all participants.

Experiment 1

Experiment 1 employed a within-subjects design. The independent variable was punishment, operationally defined as the average number of punishments participants made when playing the third-party role. The dependent variable was the two dimensions of punisher reputation (warmth and competence), measured using a 6-item Likert scale. Warmth dimension items included: "I think this member is: 1) trustworthy; 2) respectable; 3) friendly." Competence dimension items included: "I think this member: 4) can bring more benefits to the group; 5) their actions are helpful for maintaining group interests; 6) can play a leadership role in the group." Items 1-3 were adapted from Barclay (2006), and items 4-6 were adapted from Hardy and van Vugt (2006). All items used a 7-point scale, where 1 indicated "strongly disagree" and 7 indicated "strongly agree." The moderator variable was attribution of punishment motives, measured by one item: "Regarding this member' s punishment behavior, I think his/her performance is motivated by self-focused–group-focused motives." This item also used a 7-point scale, where 1 indicated "completely self-focused" (i.e., concerned with personal interests) and 7 indicated "completely group-focused" (i.e., concerned with collective interests).

Experiment 1 consisted of 12 rounds of third-party dictator games conducted via computer-based experiments using z-Tree (Fischbacher, 2007). Participants were randomly divided into 30 groups of 3 people each, with real names replaced by codes A, B, and C. During the experiment, participants were in separate cubicles and not allowed to communicate. Experimental instructions used neutral language (e.g., "deduction") instead of emotionally charged terms (e.g., "punishment"). Before the experiment began, participants were informed that they would play the roles of dictator, recipient, and third party with two other members (to avoid potential priming, the actual instructions used "Role A," "Role B," and "Role C" instead of dictator, recipient, and third party; the same applies to Experiment 2). In each round, participants randomly played one of the three roles, but across the entire experiment, each participant played each role an

equal number of times (4 times each). At the beginning of each round, the dictator received an initial endowment of 10 tokens (equivalent to 30 RMB) from the experimenter, while the third party and recipient received 5 and 0 tokens, respectively. The dictator could allocate any proportion of the amount to the recipient according to their will, and the recipient could not reject the allocation, regardless of fairness. If the third party considered the allocation unfair, they could punish the dictator. The punishment rule was uniformly set as: the third party paid 2 tokens to deduct 6 tokens from the dictator.

After the experiment began, the dictator made an allocation, the third party decided whether to punish after seeing the allocation, and then the allocation scheme and the third party' s punishment decision were presented on each participant' s screen. After the final round, the experimenter provided feedback to each participant about the other two group members' performance across the 12 rounds, including: 1) average number of punishments made as third party; 2) average amount received as recipient; 3) total tokens held at the end of the experiment. Following the 12 rounds, participants used the aforementioned scales to evaluate the other two group members on warmth, competence, and punishment motives. After completing these steps, the experimenter explained the experimental purpose and paid participants. The payment consisted of a show-up fee plus the tokens held in a randomly selected round.

2.4 Results and Discussion

Confirmatory factor analysis of the six items across the two dimensions showed that the expected two-factor model demonstrated good fit (CMIN/DF = 3.020, RMSEA = 0.048, GFI = 0.991, CFI = 0.997, NFI = 0.995, PNFI = 0.531, PGFI = 0.378), and was significantly superior ($\Delta\ ^2$/df = 366.461, p < 0.001) to the one-factor model (CMIN/DF = 43.402, RMSEA = 0.219, GFI = 0.846, CFI = 0.924, NFI = 0.922, PNFI = 0.553, PGFI = 0.362). Table 1 presents the means, standard deviations, and correlations of all variables. Differences in the four main variables—punishment, attribution, warmth, and competence—were not significant across gender (F = 0.03-1.28, p = 0.261-0.864) or major (F = 0.48-1.45, p = 0.197-0.846).

Using competence as the dependent variable, hierarchical regression was employed to test the main and interaction effects of punishment and attribution. To reduce multicollinearity, independent variables, moderator variables, and control variables were all centered. Regression results are shown in Table 2 .

In Model M1, both punishment (B = 3.52, β = 0.55, p < 0.001, 95%CI = [2.59, 4.46]) and attribution (B = 0.47, β = 0.45, p < 0.001, 95%CI = [0.32, 0.62]) had significant main effects on competence: the more punishments participants made or the more their punishment was attributed to collective focus, the higher the competence evaluation they received. These results indicate that both punishment and motivational attribution significantly affect others' evaluation of the punisher' s competence.

In Model M2, the interaction term between punishment and attribution had a significant positive effect on competence (B = 1.19, β = 0.36, p < 0.001, 95%CI = [0.75, 1.62]), explaining 12% of the variance in competence. This suggests that the effect of punishment on competence is positively moderated by attribution. To more clearly display the moderating effect of attribution, the Johnson-Neyman technique was used to further quantify how attribution affects the relationship between punishment and competence and to test the statistical significance region of the moderation effect, with results shown in Figure 1 [Figure 1: see original paper].

Figure 1 shows that when attribution exceeds 2, the confidence interval of the regression slope for punishment affecting competence is above zero, indicating that when punishment motive attribution exceeds this threshold, the more punishment is attributed to collective focus, the greater its enhancing effect on competence. When attribution is below 2, the confidence interval includes zero, and punishment' s effect on competence is not significant. These results suggest that punishment' s effect on competence is conditional: punishment attributed as self-focused is perceived as a self-interested tactic rather than norm enforcement behavior, thus unlikely to positively impact collective interests and losing its function of enhancing competence evaluation. Therefore, it is reasonable to conclude that only punishment perceived as focusing on collective interests can enhance the punisher' s competence evaluation.

Using warmth—the other dimension of reputation—as the dependent variable, the same method was used to test the main and interaction effects of punishment and attribution, with results shown in Table 3 . In Model M1, both punishment (B = -1.24, β = -0.27, p = 0.003, 95%CI = [-2.05, -0.44]) and attribution (B = 0.42, β = 0.55, p < 0.001, 95%CI = [0.29, 0.55]) had significant main effects on warmth: the more punishments participants made, the lower the warmth evaluation they received; the more punishment was attributed to collective focus, the higher the warmth evaluation. These results indicate that punishment significantly reduces warmth evaluation, while collective-focused attribution helps mitigate this negative effect.

In Model M2, the interaction term between punishment and attribution had a significant positive effect on warmth (B = 0.52, β = 0.22, p = 0.015, 95%CI = [0.10, 0.94]), explaining 4% of the variance in warmth. This indicates that the effect of punishment on warmth is moderated by attribution. The Johnson-Neyman technique was used to further quantify how attribution affects the relationship between punishment and warmth and to test the statistical significance region of the moderation effect, with results shown in Figure 2 [Figure 2: see original paper].

Figure 2 shows that when attribution is below 4.39, the confidence interval of the regression slope for punishment affecting warmth is below zero, indicating that below this threshold, the more punishment is attributed to individual focus, the greater its negative effect on warmth evaluation. When attribution exceeds 4.39, the confidence interval includes zero, and punishment' s effect on warmth

is not significant, suggesting that when motivational attribution leans toward collective focus, the negative effect of punishment on warmth disappears. These results indicate that punishment generally reduces our evaluation of punishers on the warmth dimension, but punishment attributed as collective-focused is perceived as behavior that maintains group norms and enhances group interests, thereby eliminating the negative effect on warmth. Therefore, it is reasonable to conclude that as long as punishment is considered sufficiently motivated by maintaining collective interests, it will not reduce the punisher' s warmth evaluation.

Experiment 1 Discussion

Experiment 1 provided preliminary answers to Research Questions 1 and 2. Two important conclusions can be drawn from its results: First, punishment behavior has significantly different effects on the two dimensions of punisher reputation. In short, punishment generally enhances competence evaluation while reducing warmth evaluation. This means the effects on the two reputation dimensions are opposite in direction, which partially explains seemingly contradictory findings in previous research. For example, de Kwaadsteniet et al. (2019) noted that people evaluate leaders who punish negligent employees more highly than those who never punish, yet paradoxically prefer the latter. Barclay (2006) reported similar findings. Based on Experiment 1' s results, we argue this is because punishment enhances competence evaluation while simultaneously reducing warmth evaluation. Second, attribution of punishment motives significantly affects others' evaluation of punisher reputation. Specifically, the more punishment motives are perceived as concerned with collective interests, the more they enhance punishment' s positive effect on competence and reduce its negative effect on warmth. This also suggests a potential bidirectional mechanism between punishment and social norms. While current research has focused on how third-party punishment maintains social norms (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002), some scholars have noted that punishment can negatively affect cooperation when lacking reasonable social norm guidance (Bicchieri et al., 2018; Fehr & Rockenbach, 2003). Experiment 1 indicates this effect also exists in punisher reputation—only punishment considered norm-compliant (collective-focused) can potentially enhance punisher reputation overall.

Experiment 2

Experiment 1 demonstrated that third-party punishment differentially affects the two dimensions of punisher reputation, and that attribution of punishment motives influences reputation evaluation. A remaining question concerns what cues people use to attribute punishment motives. As Kiyonari and Barclay (2008) noted, in real life, bystanders are unlikely to have complete information about the antecedents and consequences of punishment and must rely on available cues to judge punishment motives. In Experiment 2, we introduced informational cues that participants could use to judge punishment motives.

We hypothesized that the punisher' s own cooperation level would, to some extent, signal whether their punishment motives were reasonable. For instance, individuals who never contribute to public goods or are stingy when allocating resources would seem unlikely to punish for norm maintenance motives. Additionally, although financial punishment remains mainstream in laboratory third-party punishment research, Guala (2012) observed that, contrary to laboratory settings, people in real life prefer social sanctions over financial sanctions to discipline violators. Therefore, another purpose of Experiment 2 was to introduce both financial and social punishment forms and examine their effects on punisher reputation.

We used G*Power 3.1 to determine the sample size for Experiment 2: with a medium effect size of $f^2 = 0.0625$ and significance level $\alpha = 0.05$, 171 participants were needed to achieve 95% statistical power (1-β). A total of 176 social participants actually participated in Experiment 2. Participants had a mean age of 35.07 $\pm$ 17.49 years, with 59.66% being female. Occupational distribution was: students 25.57%, government and public institutions 18.75%, various enterprises 24.43%, self-employed 19.32%, and others 11.93%. Educational distribution was: secondary technical school and below 27.27%, junior college 21.59%, undergraduate 45.45%, and master' s and doctoral degrees 5.68%. Monthly income distribution was: below 2000 RMB 10.80%, 2000–5000 RMB 28.41%, 5000–10,000 RMB 44.89%, and above 10,000 RMB 15.91%. All participants had never participated in similar experiments and signed informed consent forms before the experiment began.

Experiment 2 employed a 2 (cooperation: low/high) $\times$ 2 (financial punishment: absent/present) $\times$ 2 (social punishment: absent/present) within-subjects design. Cooperation was operationally defined as the amount participants allocated to recipients when playing the dictator role. Financial punishment was operationally defined as participants paying 2 tokens to deduct 6 tokens from the dictator. Social punishment was operationally defined as participants sending the message "I think your allocation is unfair" to the dictator (Nelissen & Mulder, 2013). As in Experiment 1, the dependent variable was the two dimensions of punisher reputation, measured using the same 6-item Likert scale (see Experiment 1).

Experiment 2 also consisted of 12 rounds of third-party dictator games with procedures similar to Experiment 1, except for: 1) informing participants they would play with 8 other members in the roles of dictator, recipient, and third party, when in fact the other 8 members were not real participants but preprogrammed by the experimenter; 2) informing participants that in each round, the 9 members would be randomly divided into 3 subgroups, each containing one dictator, one recipient, and one third party. When participants played the third party, the experimenter provided feedback about their subgroup' s allocation; when participants played other roles, no feedback was provided for that round. This arrangement ensured that although each round involved 3 members, the random composition of subgroups each round gave participants equal

chances to interact directly with each of the 8 virtual members; 3) when facing an unfair allocation, the third party could choose no punishment, financial punishment, social punishment, or both; 4) in each round, participants randomly played dictator, recipient, or third party, but across the entire experiment, each participant played each role an equal number of times (4 times each); 5) after the final round, the experimenter provided feedback about the other 8 members' performance across the 12 rounds, including: allocation level to recipients as dictator (low/high); whether they had engaged in financial punishment as third party (absent/present); whether they had engaged in social punishment as third party (absent/present). In fact, the feedback was pre-set by the experimenter, containing all 8 combinations of 2 (cooperation: low/high) × 2 (financial punishment: absent/present) × 2 (social punishment: absent/present), with each combination corresponding to one member. All participants saw the same feedback but in random order. Participants then evaluated the other 8 members and attributed motives to their punishment using the same scales as in Experiment 1.

3.4 Results and Discussion

First, we tested whether cooperation level significantly affected participants' attribution of punishment motives: attribution for high-cooperation punishers ($M = 3.01$, $SD = 1.45$) was significantly higher than for low-cooperation punishers ($M = 2.45$, $SD = 1.81$) ($t = 6.46$, $p < 0.001$, $d = 0.34$, 95%CI = [0.24, 0.45]), indicating that high-cooperation punishers' motives were more likely to be attributed as collective-focused. Thus, as predicted, the punisher' s cooperation behavior indeed serves as an important attribution cue. Differences in the two main variables—warmth ($F = 0.23$–1.01, $p = 0.463$–0.921) and competence ($F = 0.60$–1.64, $p = 0.07$–0.62)—were not significant across gender, occupation, education level, or income level. Age was not significantly correlated with warmth ($r = -0.04$, $p = 0.635$) or competence ($r = 0.09$, $p = 0.247$). Table 4 presents the descriptive statistics.

A 2 × 2 × 2 multivariate analysis of variance (MANOVA) was conducted with warmth and competence as dependent variables and cooperation, social punishment, and financial punishment as independent variables. Multivariate tests showed significant main effects of cooperation (Wilks' Lambda = 0.82, $F = 157.17$, $p < 0.001$, partial $^2 = 0.18$), social punishment (Wilks' Lambda = 0.97, $F = 22.77$, $p < 0.001$, partial $^2 = 0.03$), and financial punishment (Wilks' Lambda = 0.96, $F = 29.04$, $p < 0.001$, partial $^2 = 0.04$) on the two dependent variables. Significant interaction effects were also found for cooperation × social punishment (Wilks' Lambda = 0.99, $F = 5.15$, $p = 0.006$, partial $^2 = 0.01$) and financial punishment × social punishment (Wilks' Lambda = 0.99, $F = 10.99$, $p < 0.001$, partial $^2 = 0.02$). However, cooperation × financial punishment (Wilks' Lambda = 0.99, $F = 0.88$, $p = 0.415$) and the three-way interaction (Wilks' Lambda = 1, $F = 0.08$, $p = 0.929$) were not significant. These results indicate that both social and financial punishment directly affect reputation,

and that social punishment' s effect on reputation differs across cooperation levels, while financial punishment' s effect on reputation differs across social punishment levels.

Further tests of between-subjects effects are shown in Table 5 . Cooperation had significant main effects on both warmth and competence. Financial punishment had a significant main effect on warmth but not on competence. Social punishment had a significant main effect on competence but not on warmth. Additionally, the cooperation × social punishment interaction was significant on both competence and warmth dimensions, and the financial punishment × social punishment interaction was significant on the competence dimension.

Since the cooperation × social punishment interaction was significant on both dimensions, we further analyzed the simple effects of social punishment at different cooperation levels. Multivariate tests showed significant simple effects of social punishment on both dependent variables at low cooperation (Wilks' Lambda = 0.99, F = 7.27, p = 0.001, partial ^{2} = 0.01) and high cooperation levels (Wilks' Lambda = 0.97, F = 20.65, p < 0.001, partial ^{2} = 0.03), with a larger effect size at high cooperation. Univariate tests revealed that at low cooperation, social punishment had a significant simple effect on warmth (F = 6.22, p = 0.013, partial ^{2} = 0.004) but not at high cooperation (F = 0.13, p = 0.721), indicating that social punishment by low-cooperation members significantly reduced warmth evaluation, while social punishment by high-cooperation members did not negatively affect warmth. For competence, social punishment had no significant simple effect at low cooperation (F = 0.002, p = 0.961) but a significant effect at high cooperation (F = 20.50, p < 0.001, partial ^{2} = 0.01), indicating that only social punishment by high-cooperation members significantly enhanced competence evaluation. Pairwise comparisons (Bonferroni-corrected) further confirmed these findings (see Figures 3 [Figure 3: see original paper] and 4 [Figure 4: see original paper]): at low cooperation, warmth evaluation was significantly lower for members who engaged in social punishment (M = 3.22, SE = 0.08) compared to those who did not (M = 3.48, SE = 0.08) (p = 0.013, 95%CI = [0.06, 0.47]); at high cooperation, warmth evaluation did not differ significantly between members who engaged in social punishment (M = 4.70, SE = 0.08) and those who did not (M = 4.66, SE = 0.075) (p = 0.721, 95%CI = [-0.25, 0.17]). At high cooperation, competence evaluation was significantly higher for members who engaged in social punishment (M = 4.70, SE = 0.08) compared to those who did not (M = 4.18, SE = 0.08) (p < 0.001, 95%CI = [-0.75, -0.30]); at low cooperation, competence evaluation did not differ significantly between members who engaged in social punishment (M = 3.34, SE = 0.08) and those who did not (M = 3.33, SE = 0.08) (p = 0.961, 95%CI = [-0.23, 0.22]).

Since the financial punishment × social punishment interaction was significant on the competence dimension, we further analyzed the simple effects of financial punishment at different social punishment levels. Multivariate tests showed significant simple effects of financial punishment on both dependent variables

when social punishment was absent (Wilks' Lambda = 0.96, F = 33.06, p < 0.001, partial ^{2} = 0.05) and present (Wilks' Lambda = 0.99, F = 6.97, p = 0.001, partial ^{2} = 0.01), with a larger effect size when social punishment was absent. Univariate tests revealed that financial punishment had significant simple effects on warmth both when social punishment was absent (F = 11.00, p = 0.001, partial ^{2} = 0.01) and present (F = 13.30, p < 0.001, partial ^{2} = 0.01), indicating that financial punishment significantly reduced warmth evaluation regardless of social punishment. Financial punishment also had significant simple effects on competence both when social punishment was absent (F = 6.00, p = 0.014, partial ^{2} = 0.004) and present (F = 4.81, p = 0.028, partial ^{2} = 0.003), but the direction of effect differed. Pairwise comparisons (Bonferroni-corrected) further showed (see Figures 5 [Figure 5: see original paper] and 6 [Figure 6: see original paper]): when social punishment was absent, warmth evaluation was significantly lower for members who engaged in financial punishment (M = 3.89, SE = 0.08) compared to those who did not (M = 4.25, SE = 0.08) (p = 0.001, 95%CI = [0.14, 0.56]); when social punishment was present, warmth evaluation was also significantly lower for members who engaged in financial punishment (M = 3.76, SE = 0.08) compared to those who did not (M = 4.15, SE = 0.08) (p < 0.001, 95%CI = [0.18, 0.59]). When social punishment was absent, competence evaluation was significantly higher for members who engaged in financial punishment (M = 3.90, SE = 0.081) compared to those who did not (M = 3.62, SE = 0.08) (p = 0.014, 95%CI = [-0.51, -0.06]); when social punishment was present, competence evaluation was significantly lower for members who engaged in financial punishment (M = 3.89, SE = 0.08) compared to those who did not (M = 4.15, SE = 0.08) (p = 0.028, 95%CI = [0.03, 0.48]). Thus, the direction of financial punishment' s effect on competence was opposite depending on whether social punishment was present or absent.

Experiment 2 Discussion

Experiment 1 preliminarily answered Research Question 1 regarding whether bystanders' attribution of punishment motives differentially affects the two reputation dimensions. Experiment 2 extended this by examining what cues bystanders use to attribute punishment motives. The results showed that the punisher' s own cooperation level serves as an important cue: high cooperation leads bystanders to attribute punishment motives to collective interest, while low cooperation signals self-interest. Experiment 1 demonstrated that punishment generally reduces warmth evaluation while enhancing competence evaluation, and that collective-focused attribution significantly reduces the negative effect on warmth while further enhancing the positive effect on competence. Experiment 2 observed similar results: social punishment by high-cooperation punishers did not affect warmth evaluation but enhanced competence evaluation.

Additionally, Experiment 2 partially answered Research Question 3 about whether different punishment forms have different effects on reputation. The

results affirmatively showed that financial punishment did not significantly affect competence evaluation but reduced warmth evaluation, whereas social punishment did not significantly affect warmth but enhanced competence evaluation. Furthermore, the interaction analysis supported existing literature' s conclusion that financial punishment often produces side effects (Chen & Zhu, 2020; Houser et al., 2008), especially when individuals have other options (Xie & Su, 2019). In Experiment 2, the side effect of financial punishment primarily manifested as consistently reducing warmth evaluation regardless of whether individuals engaged in social punishment, possibly because social punishment is considered a better option for maintaining social norms (Cui et al., 2017). When social punishment is available, using financial punishment may be perceived as motivated by negative intentions such as self-interest and malice (Fehr & Rockenbach, 2003), thereby reducing the punisher' s warmth evaluation. The pattern differed for competence: in the absence of social punishment, financial punishment enhanced competence evaluation, but when social punishment was present, it reduced competence evaluation. We speculate this relates to punishment effectiveness and efficiency (Balliet et al., 2011): effectiveness refers to whether punishment increases cooperation, while efficiency refers to whether punishment increases collective net benefits after deducting punishment costs. Without social punishment, financial punishment objectively signals social norms (regardless of motivation) and promotes (future) cooperative behavior among violators to some extent (Bicchieri et al., 2018; Chen et al., 2020), thus having a positive effect from an effectiveness perspective and enhancing competence evaluation. Conversely, when social punishment already signals norms, the additional effect of financial punishment may be insignificant, while its high costs may reduce collective net benefits (Dreber et al., 2008), thus having a potential negative effect from an efficiency perspective and reducing competence evaluation.

Experiment 3

Experiments 1 and 2 essentially answered the three research questions posed in this paper. However, to further examine the relationships among the main variables, we still need to test the interaction mechanism between attribution of punishment motives and punishment form. Experiment 3 aims to address this issue. Additionally, while the first two experiments examined how punishers' performance in other roles affected their reputation (Experiment 1 showed that performance as a recipient did not affect reputation as a punisher, while Experiment 2 showed that performance as a dictator significantly affected reputation as a punisher), Experiment 3 modified the classic dictator game paradigm. We no longer examined role effects but focused on the interaction between punishment form and attribution. Finally, since the first two experiments had already examined the effect of whether to punish on reputation, Experiment 3 no longer included a "no punishment" option.

4.1 Participants

We used G*Power 3.1 to determine the required sample size: with a medium effect size of $f^2 = 0.15$ and significance level $\alpha = 0.05$, 119 participants were needed to achieve 95% statistical power (1-β). A total of 120 undergraduate students from a university who were not psychology majors actually participated in Experiment 3. Participants had a mean age of 21.20 ± 1.72 years, with 53.33% being female. Major distribution was: science and engineering 32.50%, social sciences 29.17%, humanities 20.83%, and arts and other fields 17.50%.

Experiment 3 employed a within-subjects design. The independent variable was punishment form (financial punishment vs. social punishment, operationally defined as in Experiment 2). The dependent variable was the two dimensions of punisher reputation, measured using the same scale as in Experiment 1. The moderator variable was attribution of punishment motives, measured by the question: "Regarding this member' s punishment behavior, I think his/her performance is motivated by concern for self-interest (self-focused)—concern for collective interest (group-focused)." This question used a 7-point scale, where 1 indicated "completely self-focused" and 7 indicated "completely group-focused."

Experiment 3 used a dictator game with multiple third parties (Ouss & Peysakhovich, 2015), where each group consisted of 1 dictator (Role A), 1 recipient (Role B), and 2 third parties (Roles C and D). Before the game, the dictator, recipient, and two third parties had initial endowments of 10, 0, and 5 tokens, respectively. The dictator could freely allocate the initial amount between themselves and the recipient, who had no veto power, but both third parties could punish allocations they deemed unfair. Punishment had two levels: financial and social. In Experiment 3, all participants were bystanders who did not directly participate in the game but observed one round of the dictator game with 4 players. Their task was to calculate each player' s final earnings as quickly as possible after the game concluded. The 4 players were actually pre-programmed by the experimenter. Participants observed an allocation where the dictator gave 2 tokens (20% of the initial amount) to the recipient, and Third Party C and Third Party D engaged in financial and social punishment, respectively. Participants then calculated each individual' s earnings, evaluated the dictator and two third parties, and attributed motives to the two third parties' punishment.

4.4 Results and Discussion

In Experiment 3, differences in the three main variables—attribution, warmth, and competence—were not significant across gender ($F = 1.09$–1.47, $p = 0.23$–0.30) or major ($F = 1.38$–1.42, $p = 0.238$–0.740). Descriptive statistics for the three variables are shown in Table 6 .

A hierarchical regression was conducted with punishment form (financial punishment = 1, social punishment = 0) as the independent variable, attribution as the moderator, and warmth as the dependent variable to test the moderating

effect of attribution. Results are shown in Table 7 . In Model M1, both punishment form (B = 0.80, $\beta = 0.28$, p < 0.001, 95%CI = [0.45, 1.14]) and attribution (B = 0.12, $\beta = 0.14$, p = 0.027, 95%CI = [0.01, 0.23]) had significant main effects, indicating that both punishment form and attribution significantly affect warmth evaluation. In Model M2, the interaction between punishment form and attribution was also significant (B = -0.30, β = -0.47, p = 0.005, 95%CI = [-0.52, -0.09]), with a significant increase in R^2 change, explaining 3% of the variance in warmth. The negative coefficient of the interaction term indicates that as participants increasingly attribute punishment to collective focus, the effect of punishment form on warmth gradually diminishes. To more clearly display the moderating effect of attribution, the Johnson-Neyman technique was used to further quantify how attribution affects the relationship between punishment form and warmth and to test the statistical significance region of the moderation effect, with results shown in Figure 7 [Figure 7: see original paper].

Figure 7 shows that when attribution is below 4.89, the confidence interval of the regression slope for punishment form affecting warmth is above zero, indicating that punishment form significantly affects warmth, with participants giving higher warmth evaluations to individuals who engaged in social punishment compared to financial punishment. When attribution exceeds 4.89, the confidence interval includes zero, and punishment form no longer significantly affects warmth, with no significant difference in warmth evaluations between the two types of punishers. These results indicate that the effect of punishment form on warmth is conditional: when punishment is attributed as self-focused, participants give lower warmth evaluations to individuals who engaged in financial punishment; when punishment is attributed as collective-focused, warmth evaluations do not differ significantly between the two punishment types.

A further hierarchical regression was conducted with punishment form (financial punishment = 1, social punishment = 0) as the independent variable, attribution as the moderator, and competence as the dependent variable. Results are shown in Table 8 . In Model M1, both punishment form (B = 0.70, $\beta = 0.22$, p < 0.001, 95%CI = [0.31, 1.10]) and attribution (B = 0.26, $\beta = 0.25$, p < 0.001, 95%CI = [0.13, 0.38]) had significant main effects, indicating that both punishment form and attribution significantly affect competence evaluation. In Model M2, the interaction between punishment form and attribution was also significant (B = 0.51, $\beta = 0.69$, p < 0.001, 95%CI = [0.28, 0.75]), with a significant increase in R^2 change, explaining 6% of the variance in competence. The positive coefficient of the interaction term indicates that as participants increasingly attribute punishment to collective focus, the effect of punishment form on competence gradually increases. To more clearly display the moderating effect of attribution, the Johnson-Neyman technique was used to further quantify how attribution affects the relationship between punishment form and competence, with results shown in Figure 8 [Figure 8: see original paper].

Figure 8 shows that when attribution exceeds 3.29, the confidence interval of the regression slope for punishment form affecting competence is above zero, indi-

cating that punishment form significantly affects competence, with participants giving higher competence evaluations to individuals who engaged in social punishment compared to financial punishment. When attribution is below 3.29, the confidence interval includes zero, and punishment form does not significantly affect competence, with no significant difference in competence evaluations between the two types of punishers. These results indicate that the effect of punishment form on competence is conditional: when punishment is attributed as collective-focused, participants give higher competence evaluations to individuals who engaged in social punishment; when punishment is attributed as self-focused, competence evaluations do not differ significantly between the two punishment types.

Although financial punishment remains the mainstream paradigm in laboratory third-party punishment research, some scholars have begun exploring the role of social punishment in cooperation (Noussair & Tucker, 2005). Nelissen and Mulder (2013) compared social and financial punishment in promoting cooperation and found the former more effective, while Wu et al. (2016) noted that social punishment not only more effectively promoted cooperation among individuals but also increased collective net benefits. Experiment 3 compared the effects of the two punishment forms on reputation and found that social punishment' s superiority over financial punishment also extends to punisher reputation: when punishment motives are attributed as individual-focused, financial punishment has a significantly stronger negative effect on warmth than social punishment; when attributed as collective-focused, financial punishment has a significantly weaker positive effect on competence than social punishment. Previous research on punisher reputation has largely been based on financial punishment (e.g., Barclay, 2006; Hardy & van Vugt, 2006; Kiyonari & Barclay, 2008). Our findings suggest that introducing different punishment forms provides new insights for punisher reputation research. Additionally, unlike the first two experiments where participants directly participated in the game, participants in Experiment 3 were bystanders without direct interaction with punishers, eliminating role effects (where participants' performance as dictators or recipients affects their reputation as punishers). Nevertheless, we obtained similar results, suggesting our conclusions have high robustness.

5. General Discussion

In second-order social dilemmas, third-party punishment itself is a public good (Colman, 2006; Hauert et al., 2007). How the provider of this public good—the punisher—emerges and is selected in evolution is a challenge for researchers. An intuitive answer is that third-party punishment brings positive reputation and resulting additional benefits to punishers, which can offset punishment costs in the long term (Barclay, 2006; Barclay & Kiyonari, 2014), conferring evolutionary advantages. The premise of this theory is that the reputation derived from third-party punishment must be positive, yet existing literature suggests this assumption may not hold (de Kwaadsteniet et al., 2019; Ozono & Watabe, 2012).

This paper explores the psychological mechanisms underlying this phenomenon from the perspectives of reputation' s dual dimensions, punishment motives, and punishment forms, advancing our understanding of punisher reputation mechanisms in several ways.

First, previous research has tended to view reputation as a uni-dimensional variable (de Kwaadsteniet et al., 2019), with punishment' s effect on reputation being unidirectional—either positive (Barclay, 2006; Barclay & Kiyonari, 2014) or negative (Ozono & Watabe, 2012). Based on Fiske et al.' s (2007) theory, this study divides reputation into warmth and competence dimensions. Results show that punishment' s effects on these two dimensions are opposite: punishment reduces warmth evaluation while enhancing competence evaluation. Using Chinese classical literature' s distinction, punishers seem closer to "the wise" than "the benevolent." This implies that if future researchers attempt to use reputation to explain the evolution of third-party punishment, they must distinguish between these dimensions. In different contexts, third-party punishment brings 截然不同的 consequences for punishers. If a group facing crisis prefers members with outstanding competence, punishers may gain higher power or social status (Gross et al., 2016) because their punishment behavior results in higher competence evaluation. However, if the group prefers friendly and gentle members for various reasons, punishers may face negative consequences such as exclusion or reduced probability of receiving help due to their lower warmth evaluation, meaning they are disliked (Geiger & Swim, 2016). In short, reputation mechanisms can only partially explain punishers' selective advantages in specific situations, so other mechanisms must contribute to the selection and diffusion of third-party punishment in evolution (Dreber et al., 2008). Exploring these potential mechanisms is an important direction for future research.

Second, influenced by economics and biology, motivation has been largely neglected in third-party punishment literature, with punishment' s effect on punisher reputation understood as a simple "stimulus-response" behaviorist model: punishment directly triggers positive or negative evaluations without considering the subjective motives driving it. However, life experience and psychological literature indicate that interpersonal interactions largely depend on inferences about participants' behavioral motives (Bigman & Tamir, 2016). By introducing a motivational perspective, this study confirms that motivational attribution affects punisher reputation. Specifically, punishment attributed as collective-focused mitigates its negative effect on warmth and enhances its positive effect on competence, whereas punishment attributed as self-focused further reduces warmth evaluation and loses its positive function of enhancing competence evaluation. This finding means that the same punishment behavior has 截然不同的 effects on punisher reputation under different motivational attributions. Therefore, incorporating punishment motive attribution is crucial for understanding punisher reputation mechanisms. This result also partially explains why reputation mechanisms cannot fully account for punishers' selective advantages. Chen and Yang' s (in press) analysis of third-party punishment motives shows that punishers' behavior is largely driven by self-interested motives, which 反而 hinder

punishers from obtaining good reputation.

Third, Guala (2012) pointed out that real-life third-party punishment more often takes the form of social punishment rather than the financial punishment common in laboratory settings. This study enhances the applicability of social punishment by introducing different punishment forms and examining their interaction with punishment motives. Previous research suggests social punishment is more effective than financial punishment in promoting cooperation and increasing collective benefits (Nelissen & Mulder, 2013; Wu et al., 2016). Our findings show this effect also extends to punisher reputation: financial punishment generally reduces punisher reputation while social punishment generally enhances it, and punishment motive attribution amplifies these effects: self-focused attribution further magnifies financial punishment' s negative effect on warmth, while collective-focused attribution further enhances social punishment' s positive effect on competence. Regarding the interaction between the two punishment forms, we found that when social punishment is available, financial punishment always reduces warmth evaluation regardless of whether individuals engage in social punishment. For competence, financial punishment alone enhances reputation, but combining both forms 反而 negatively affects reputation. This suggests that punishment as a means of maintaining social norms is not "the more, the better." This 呼应 s scholars' mentions of punishment' s potential negative effects (Chen et al., 2020; Fehr & Williams, 2018) and has practical implications for policymakers: excessive punishment may reduce punisher reputation and decrease social efficiency.

Fourth, this study' s theoretical contribution lies in its preliminary exploration of cues individuals use to attribute punishment motives. As Kiyonari and Barclay (2008) noted, people in real life cannot fully track the antecedents and consequences of punishment, so they must rely on limited cues to infer punishers' motives. Our results preliminarily indicate that punishers' cooperation level serves as such a cue: high cooperation implies punishment motives are for maintaining collective interests. Additionally, previous literature suggests social punishment generally has positive effects on cooperation (Cui et al., 2017; Nelissen & Mulder, 2013). Our findings indicate this may be because previous studies lacked other signaling mechanisms, so social punishment was always perceived as well-intentioned. However, when social punishment contradicts other signals (e.g., low cooperation level), it can negatively affect reputation, such as reducing warmth evaluation. This leads us to question whether social punishment still positively affects cooperation under such circumstances. This finding' s significance lies in that punishment' s effect on reputation or cooperation does not exist in a vacuum; rather, it is rooted in the punisher' s various behaviors, including both punishment and non-punishment behaviors. Among the latter, some behaviors (e.g., cooperation) serve as cues for inferring punishment motives, while others (e.g., performance as a recipient) lack this function. Most laboratory studies have artificially eliminated these cues to obtain clearer causal relationships, but our results suggest conclusions obtained this way may be somewhat one-sided. Future research could improve study design in two ways:

1) incorporate relevant signaling cues and examine their effects on punishment; 2) test additional signaling cues and their interactions.

Finally, despite yielding several meaningful results, this exploratory study on the relationships among punishment motives, punishment forms, and different dimensions of punisher reputation has various limitations. First, regarding punishment forms, an issue worth further discussion is the conversion relationship between financial and social punishment—that is, how many units of financial punishment intensity are equivalent to corresponding units of social punishment. Solving this would allow us to compare the two punishment forms while controlling for punishment intensity, greatly enhancing the persuasiveness of research conclusions. Second, although this study preliminarily examined cooperation as an attribution cue, real-life cues suggesting punishment motives are obviously much richer. Therefore, besides punishers' cooperation level, we need to explore other cues and their interactions, which we could not analyze further due to research technology and article length constraints. Finally, in Experiments 1 and 2, evaluators directly interacted with punishers, whereas in Experiment 3, evaluators did not directly interact with punishers but participated as bystanders. These correspond to two typical real-life situations (whether evaluators directly participated in the event they evaluate), but we did not directly compare the effects of these two conditions on punisher reputation. Future research could further examine the effect of whether evaluators participate in the game, which has important practical significance because in real life, we may either directly participate in an event or merely be bystanders when evaluating others.

References

Balliet, D., Mulder, L. B., & van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594–630.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325–337.

Barclay, P., & Kiyonari, T. (2014). Why sanction? Functional causes of punishment and reward. In P. A. van Lange, B. Rockenbach, & T. Yamagishi (Eds.), *Reward and punishment in social dilemmas* (pp. 182–196). Oxford, England: Oxford University Press.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences of the United States of America, 113*(4), 936–941.

Beersma, B., & van Kleef, G. A. (2011). How the grapevine keeps you in line: Gossip increases contributions to the group. *Social Psychological and Personality Science, 2*, 642–649.

Bicchieri, C., Dimant, E., & Xiao, E. T. (2018). Deviant or wrong? The effects of norm information on the efficacy of punishment (PPE Working Papers

0016). Philadelphia, PA: Philosophy, Politics and Economics of University of Pennsylvania.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General, 145*(12), 1654–1669.

Bornstein, G., & Weisel, O. (2010). Punishment, cooperation, and cheater detection in "noisy" social exchange. *Games, 1*(1), 18–33.

Carpenter, J., Bowles, S., Gintis, H., & Hwang, S. H. (2009). Strong reciprocity and team production: Theory and evidence. *Journal of Economic Behavior & Organization, 71*(2), 221–232.

Chen, H., Zeng, Z., & Ma, J. (2020). The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *PloS One, 15*(3), e0229510.

Chen, S. J., Hu, H. M., & Yang, S. S. (2020). Payment vs. retaliation: Impact of cost form on third-party punishment. *Journal of Psychological Science, 43*(2), 416–422.

Chen, S. J., & Yang, S. S. (in press). Motives of altruistic punishment. *Advances in Psychological Science*.

Chen, S. J., & Zhu, Y. (2020). The other face of punishment: Detrimental effects of punishment and destructive punishment. *Journal of Psychological Science, 43*(4), 911–917.

Chen, X., Zhao, G. X., & Ye, H. S. (2014). The forms and functions of punishment in public-goods dilemmas. *Advances in Psychological Science, 22*(1), 160–170.

Colman, A. M. (2006). The puzzle of cooperation. *Nature, 440*(7088), 744–745.

Cui, L. Y., He, X., Luo, J. L., Huang, X. J., Cao, W. J., & Chen, X. M. (2017). The effects of moral punishment and relationship punishment on junior middle school students' cooperation behaviors in public goods dilemma. *Acta Psychologica Sinica, 49*(10), 1322–1333.

de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology, 84*, 103800.

de Kwaadsteniet, E. W., van Dijk, E., Wit, A., de Cremer, D., & de Rooij, M. (2007). Justifying decisions in social dilemmas: Justification pressures and tacit coordination under environmental uncertainty. *Personality and Social Psychology Bulletin, 33*(12), 1648–1660.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don' t punish. *Nature, 452*(7185), 348-351.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137-140.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*(6928), 137-140.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour, 2*, 429-430.

Fehr, E., & Williams, T. (2018). Social norms, endogenous sorting and the culture of cooperation. (ECON Working Papers 267). Zurich, Switzerland: Department of Economics of University of Zurich.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*(2), 171-178.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83.

Geiger, N., & Swim, J. K. (2016). Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. *Journal of Environmental Psychology, 47*, 79-90.

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan: Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports, 6*, 20767.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35*(1), 1-15.

Hardy, C. L., & van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin, 32*(10), 1402-1413.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science, 316*(5833), 1905-1907.

Houser, D., Xiao, E. T., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior, 62*, 509-532.

Hu, T. Y., Li, J., Jia, H., & Xie, X. (2016). Helping others, warming yourself: Altruistic behaviors increase warmth feelings of the ambient environment. *Frontiers in psychology, 7*, 1349.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473–476.

Jordan, J. J., & Rand, D. G. (2019). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology, 118*(1), 153–169.

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology, 95*(4), 826–842.

Kurzban, R., DeScioli, P., & O' Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28*(2), 75–84.

Liu, G. F., & Xin, Z. Q. (2014). The effects of punishment impacting on social trust and cooperation: Controversy and interpretation. *Journal of Shanghai Normal University (Philosophy & Social Sciences Edition), 43*(1), 146–152.

Nelissen, R. M., & Mulder, L. B. (2013). What makes a sanction "stick" ? The effects of financial and social sanctions on norm compliance. *Social Influence, 8*(1), 70–80.

Noussair, C., & Tucker, S. (2005). Combining monetary and social sanctions to promote cooperation. *Economic Inquiry, 3*(3), 649–660.

Ouss, A., & Peysakhovich, A. (2015). When punishment doesn' t pay: Cold glow and decisions to punish. *The Journal of Law and Economics, 58*(3), 625–655.

Ozono, H., & Watabe, M. (2012). Reputational benefit of punishment: Comparison among the punisher, rewarder, and non-sanctioner. *Letters on Evolutionary Behavioral Science, 3*(2), 21–24.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science, 27*(9), 1192–1206.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences, 17*(8), 413–425.

Santos, M. D., Rankin, D. J., & Wedekind, C. (2010). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences, 278*(1704), 371–377.

Wu, J., Balliet, D., & van Lange, P. A. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports, 6*, 23919.

Xie, D. J., & Su, Y. J. (2019). The evolutionary and cognitive mechanisms of third-party punishment. *Journal of Psychological Science, 42*(1), 216–222.

Xie, X. F., Wang, Y. L., Gu, S. Y., & Li, W. (2017). Is altruism just other-benefiting? A dual pathway model from an evolutionary perspective. *Advances*

in Psychological Science, 25(9), 1441–1455.

Appendix: Test Questions for Understanding Experimental Procedures and Concepts (Examples)

1. Suppose in one round, A allocates 2 tokens to B, and C does not deduct any tokens from A. The final token amounts for the three would be: (d)
   a) A: 10, B: 0, C: 0
   b) A: 8, B: 2, C: 0
   c) A: 2, B: 8, C: 5
   d) A: 8, B: 2, C: 5
2. Suppose X' s decisions mainly focus on how many tokens they receive, while Y' s decisions mainly focus on whether everyone can obtain good benefits. X and Y respectively belong to: (a)
   a) Self-focused; Group-focused
   b) Group-focused; Self-focused
   c) Self-focused; Self-focused
   d) Group-focused; Group-focused
3. In the self-focused–group-focused question, a higher score indicates: (c)
   a) More concern for individual interests
   b) More concern for others' interests
   c) More concern for collective interests
   d) More concern for third-party interests
4. When rating a statement, a lower score represents: (b)
   a) More agreement with the statement
   b) More disagreement with the statement
   c) More ignoring of the statement
   d) Considering the statement less important
5. Suppose in one round, A allocates 3 tokens to B, and C deducts tokens from A. The final token amounts for the three would be: (b)
   a) A: 1, B: 3, C: 0
   b) A: 1, B: 3, C: 3

c) A: 7, B: 3, C: 5

d) A: 1, B: 3, C: 5

Note: Figure translations are in progress. See original paper for figures.