Running Head: Punishment and trust

The Dynamics of Punishment and Trust**

Long Wang
Department of Management
College of Business
City University of Hong Kong
83 Tat Chee Ave, Kowloon, Hong Kong SAR, China
longwang@cityu.edu.hk
Phone: 00852-3442-7181; fax: 00852-3442-0309

J. Keith Murnighan*
Kellogg School of Management
Northwestern University
2001 Sheridan Rd
Evanston IL 60201

Acknowledgement: This paper benefitted immeasurably from the comments and suggestions of Tim Judge and Jeanne M. Brett. The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No: CityU 11505115). An earlier version of this paper was presented and won the Best Paper Award in Contribution to Theory at the 2017 Northeast Decision Sciences Institute (NEDSI) Conference in Springfield, Massachusetts.

* Keith Murnighan passed away before Long Wang completed the final version of the manuscript submitted for publication. With a heavy heart, Long Wang publishes the article in Keith’s memory. All the faults and problems that might remain are solely Long Wang’s.

** ©American Psychological Association, 2017. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at:
http://dx.doi.org/10.1037/apl0000178
The Dynamics of Punishment and Trust

Abstract

The trade-off between mercy and justice is a classic moral dilemma, particularly for organizational leaders and managers. In three complementary studies, we investigated how resolving the ‘punishment dilemma’ influences interpersonal trust. Study 1 used controlled scenarios to show that uninvolved observers trusted leaders who administered large or medium punishment more than leaders who administered no punishment when transgressors deserved punishment. At the same time, large punishment decreased trust more than medium or no punishment for less deserving targets. Study 2’s similar scenarios showed that leaders who administered punishment lost trust when they subsequently received benefits even though it was not clear whether their benefits resulted from their act of punishment. Study 3 provided a behavioral replication of these results. These findings suggest that people trusted punishers more than non-punishers, but only when punishers’ motives were not personal revenge. In the discussion, we explore the practical and theoretical implications of these results for organizations.

Keywords: Punishment dilemma, trust, deservingness and severity of punishment, integrity, benevolence, ethics.
The Dynamics of Punishment and Trust

“current and former Amazonians...described how they tried to reconcile the sometimes-punishing aspects of their workplace.”
New York Times, August 15 2015

“Our tolerance for any such lack of empathy needs to be zero.”
Jeff Bezos, Founder & CEO of Amazon, August 17 2015

On August 15, 2015, the New York Times published a controversial article, “Inside Amazon: Wrestling Big Ideas in a Bruising Workplace,” that described a punishing work culture at Amazon (Kantor & Streitfeld, 2015). Jeff Bezos, its founder and CEO, quickly responded, indicating that Amazon would not tolerate any “callous management practices” (Streitfeld & Kantor, 2015). This example raises a variety of issues, but at its heart it highlights the notion that punishment is a highly sensitive but common reality in organizations. Employee discipline not only challenges managers and HR professionals, impacting everyday organizational action; it also has become a frequently litigated issue (Walkowiak, 2008).

Punishment is defined as a penalty inflicted for an offense, i.e., a violation of rules, norms, and expectations in a variety of organizational and social contexts (Arvey & Ivancevich, 1980; Balliet & van Lange, 2013; Becker, 1968; Boyd, et al., 2003; Fehr & Gächter, 2002; Treviño, 1992; Sigmund, Hauert, & Nowak, 2001). Punishments often get considerable attention, within and outside organizations, that creates important reputations for leaders and managers. For example, Mary Barra, the CEO of General Motors, decided to terminate 15 employees and discipline another five after an internal investigation of the company’s 11-year failure to recall vehicles equipped with a defective ignition switch (Bennett & Ramsey, 2014). In contrast, Robert Murray, an advocate for Mitt Romney and the CEO of Murray Energy, the largest privately
owned coal mining company in the U.S., fired 156 of his employees for Obama’s reelection the day after the election (Decker, 2012). Clearly, punishments like these not only affect the people who are punished; they can also create emotional and behavioral repercussions for less involved observers both within and outside their organizations.

Organizations, groups, and leaders of all kinds use rewards and punishments (“carrots and sticks”) to motivate performance and deter problematic behavior. Although punishment can have negative emotional and behavioral consequences (Luthans & Kreitner, 1985; Sims, 1979), employees tend to respond to contingent punishment with increased job satisfaction and commitment (Podsakoff & Todor, 1985; Podsakoff, et al., 2006). In addition, research has shown that altruistic punishment, i.e., costly punishment of others’ opportunistic behaviors, encourages social cooperation (Boyd, et al., 2003; Fehr & Gächter, 2002), especially in social dilemmas (Balliet, Mulder, & Van Lange, 2011; Van Lange, Rockenbach, & Yamagishi, 2014).

Punishment also affects general observers who may or may not have any direct relationships with the punishers or the punished people. Treviño (1992), for instance, noted that observers naturally evaluate the appropriateness of leaders’ punishments. Their judgements of the distributive and procedural justice of the punishment can have a host of affective, attitudinal, and behavioral consequences in organizations (Arvey & Ivancevich, 1980). Like previous research (Niehoff et al., 1998; O’Reilly & Puffer, 1989; Treviño, 1992; Treviño & Ball, 1992), we designed the current research to assess the effects of punishment on general observers rather than on punishers and their targets, for two primary reasons: first, because the literature on punishment has already documented many of the effects of punishment on punishers and their targets (e.g., see Balliet, Mulder, & Van Lange, 2011; Treviño & Weaver, 1998; Van Lange, Rockenbach, & Yamagishi, 2014; for reviews); and second, because some of these findings are
fairly straightforward and expected, e.g., egocentric action by both punishers – including defensiveness and restraint – and their targets – including anger, resentment, avoidance, and retaliation (Hermann, Thöni, & Gächter, 2008; Kiyonari & Barclay, 2008; Luthans, 1995; Luthans & Kreitner, 1985; Nikiforakis, 2008; Sims, 1979).

Our focus on observers also allowed us to more cleanly assess the effects of punishment on people’s trust of a leader-punisher. We based our research on two fundamental assumptions, that leaders and managers cannot be effective if they let inappropriate, harmful behavior go unpunished and, simultaneously, they cannot be effective if their organizational members do not trust them. Punishment, however, can present a thorny dilemma, pitting mercy against justice (or benevolence against integrity). On the one hand, punishment restores justice and confirms the integrity of a leader and the organization’s system of justice. By not letting transgressions go unpunished (or under-punished), leaders may also appear fair, just, and strong. On the other hand, punishments can create a cruel and harsh atmosphere, one that shuns mercy, forgiveness, sympathy, and humanity.

Thus, in the current research, we investigated how different resolutions of the “punishment dilemma” can influence an important outcome – interpersonal trust. In particular, we investigated a host of important contextual factors, including deservingness and the magnitude of punishment, to examine when and how punishment enhances or damages trust. The effective handling of punishment is crucial for organizational leaders: our research contributes to the literatures of trust, punishment, and leadership, as well as their practical applications, by enhancing our understanding of their dynamic relationships in both interpersonal and organizational situations.

**Trust, distrust and punishment**
Trust is commonly defined as a willingness to be vulnerable because of positive expectations of a trustees’ behavior (Colquitt, et al., 2012; Mayer et al., 1995; Rousseau et al., 1998). Distrust, in contrast, often refers to a lack of confidence in another person (Grover, 1994) or confident, negative expectations of another’s behavior, especially undesirable and even sinister behavior (Deutsch, 1960; Lewicki, McAllister, & Bies, 1998). Lewicki, McAllister, and Bies (1998) suggested that trust and distrust are not necessarily the opposite ends of a unidimensional construct although high levels of trust often create low levels of distrusts and vice-versa. Mayer et al.’s (1995) trust model suggests that ability, benevolence, and integrity are critical attributes that people rely on to determine whether a person is trustworthy enough to warrant their own vulnerable, trusting action. People can simultaneously display both trust and distrust when a person manifests trustworthiness that varies cross-situationally (Guo, Lumineau, & Lewicki, 2015; Lewicki, et al., 1998; Lewicki, Tomlinson, & Gillespie, 2006; for reviews). For example, competent people might be trusted for their ability but be distrusted for their moral character. In terms of the Mayer et al. model, although benevolence and integrity often coincide, as good deeds signal both high integrity and benevolent considerations of others’ needs and rights, they can conceptually diverge when just or ethical actions lead to less-than-benevolent outcomes. This is exactly the situation when leaders face the punishment dilemma: justice preserves integrity but not benevolence while mercy preserves benevolence but not integrity. How leaders resolve this sticky dilemma raises important questions about the importance of the virtues of integrity and benevolence in the process of creating trust: when leaders display one virtue but not the other, how will relatively independent observers react?

Costly signaling theory suggests that altruistic punishment is likely to enhance punishers’ reputation, signaling that they are fair and trustworthy (Barclay, 2006; Smith & Bliege Bird,
The empirical evidence on this issue, however, is mixed. Nelissen (2007) found that, in the context of a dictator game, people tended to trust third-party punishers more than non-punishers. Research on social dilemmas, in contrast, has shown that, even though people who engage in costly, altruistic punishment often enhance collective cooperation, people do not necessarily like, trust, or reward these punishers. Instead, punishers are often rated as obnoxious and untrustworthy (Cinyabuguma, et al., 2006; Horita, 2010; Kiyonari & Barclay, 2008; Strimling & Eriksson, 2014). In a different context, Podsakoff et al.’s (2006) meta-analysis of leader-employee relationships found that contingent punishments were positively related and non-contingent punishments were negatively related to trust in managers or supervisors. Their analysis, however, did not focus primarily on trust and it was based on only two correlations.

Thus, previous studies provide conflicting insights into the relationship between punishment and trust, and they do not address the tradeoff between integrity and benevolence in the punishment dilemma. In addition, their broad distinctions between punishment and non-punishment (or contingent and non-contingent punishment) may not have captured all of the nuances of punishment in different organizational and social contexts. For instance, potentially important factors such as the severity of a punishment, deservingness, and punishers’ perceived motives have not been central parts of past investigations. Thus, the current research extends past research by addressing these important issues. Specifically, we investigated the effects of the punishment dilemma, which asks leaders, transgressors, and observers to trade-off benevolence and integrity in their organizational and interpersonal actions and evaluations. We also investigated several key contextual factors, including deservingness and the magnitude of punishment, to build a broad, nuanced understanding of the dynamics of punishment and trust.

**The punishment dilemma**
Punishment is commonly used to regulate both human and animal behavior (Clutton-Brock & Parker, 1995). Children, for instance, learn to avoid acting inappropriately by watching and experiencing different kinds of reprimands and punishments, and organizations punish people who exhibit unwanted, undesirable behavior as a means of maintaining the integrity and the force of their rules and norms.

When people or organizations punish, they can use either a backward- or a forward-looking rationale to justify their actions (Carlsmith, Darley, & Robinson, 2002). Just deserts theory (Kant, 1797/1952), for instance, looks backward and suggests that perpetrators deserve punishment that is proportional to the severity of their actions. Thus, murder merits stronger punishments than burglary, and burglary merits stronger punishments than jaywalking, with all of these punishments intended to reinforce and maintain a balanced system of justice.

Utilitarian theory (Bentham, 1962), in contrast, takes a forward-looking approach, proposing that transgressors should be punished to prevent future transgressions. This view suggests that severe punishments act as a deterrent, reducing the likelihood of similar, future transgressions by either the transgressors or observers (Nagin, 1998; Zimring & Hawkins, 1995).

From either perspective, punishments serve the interests of organizations and societies by either retrospectively restoring the scales of justice or decreasing the likelihood of future transgressions (Carlsmith, Darley, & Robinson, 2002; Carlsmith, 2006). People who administer punishment, however, inevitably face the necessary but difficult task of doing harm in the pursuit of good (Molinsky & Margolis, 1995). This right-right moral dilemma (Kidder, 2009) pits four universal moral values against each other, i.e., justice and fairness against mercy and forgiveness (Logan & Gaes, 1993). The tension between these similarly noble moral values can make the act of punishing a particularly difficult choice, making it both less likely (Wang, Galinsky, &
Murnighan, 2009) and less effective (Luthans & Kreitner, 1985). Podsakoff et al.’s (2006) meta-analysis, for example, showed that rewards, especially contingent rewards, had much stronger positive effects on employees’ attitudes and behaviors than punishments did, with punishments leading to both positive and negative behavioral outcomes. Thus, it is far easier for leaders to dispense rewards than it is for them to administer punishments (Wang et al., 2009).

As noted, we focus on the effects of punishment and on its influence on general observers’ perceptions of a leader’s trustworthiness (Treviño, 1992; Treviño & Ball, 1992). We suggest that a leader’s punishment dilemma can also create a trust dilemma for observers because it pits integrity against benevolence. When punishers must choose between two valid moral values, observers make important judgments of punishers’ intentions and how they honor or respect these important values. These judgments can affect observers’ trust of the leader and have important organizational consequences. How observers evaluate these conflicting values may also depend on several important contextual factors.

**Deservingness and the magnitude of punishment**

Two particularly important contextual factors that can influence observers’ trust of punishers are the deservingness of transgressors and the magnitude of the punishment that they experience. Just deserts theory, for instance, suggests that punishment should be highly sensitive to contextual factors: people expect that punishments will be proportional to the harm committed. They also expect that mitigating or exacerbating circumstances will contribute to a determination of the appropriateness of a punishment’s magnitude. Inadvertent, unintended acts, for instance, typically warrant less punishment than intended transgressions, even when they both cause the same harm (Carlsmith, 2006; Carlsmith et al., 2002).
Just-deserts theory also suggests that observers’ trust of punishing leaders will depend on the interaction between transgressors’ deservingness and the enforcement of punishment. When a wrongdoing deserves to be punished, punishment should occur, and punishers should be seen as being trustworthy and having integrity. When a wrongdoing does not deserve to be punished because it was unintentional or accidental, punishment should not occur, and punishers should be seen as non-benevolent and untrustworthy. Thus, punishment behavior should generate trust in direct relation to deservingness. Whereas punishing deserving transgressors will increase trust, punishing inadvertent wrongdoers is likely to decrease it.

Hypothesis 1(H1): The relationship between punishment and trust will be moderated by perceptions of punishment deservingness. The nature of this interaction is such that punishing a transgressor who deserves punishment will increase observers’ trust of the punisher but punishing a transgressor who does not deserve punishment will decrease observers’ trust of the punisher.

In addition, both just deserts theory and related empirical results also suggest that deservingness should be related to a punishment’s magnitude (Carlsmith et al., 2002). In particular, severely damaging, intentional acts tend to be seen as deserving considerable punishment, and mildly serious, unintentional acts tend to be seen as deserving much less punishment – possibly none. Thus, over- and under-punishment that is inconsistent with the severity of a transgression should be seen as inappropriate, thereby reducing trust.

Hypothesis 2(H2): The relationship between punishment and trust will be moderated by punishment severity and perceptions of deservingness. The nature of the interaction is such that either severe punishment for less deserving transgressors or light punishment for more deserving transgressors will reduce observers’ trust of punishers.
Inappropriate punishment accentuates the punishment dilemma because a mismatch between the severity of the wrongdoing and the subsequent punishment often makes observers question a leader-punisher’s integrity or benevolence. When severe punishment is appropriate, observers should naturally question the leader’s integrity and trust them less if they under-punish. In contrast, when transgressors are punished more than they deserve, observers should question the leader’s benevolence because over-punishment suggests the presence of callous and even malevolent motives, and little basis for trust. Thus,

*Hypothesis 3 (H3): Punishing a less deserving transgressor will decrease trust with observers’ perceptions of the punisher’s benevolence mediating the effects of over-punishment on trust; punishing a more deserving transgressor will increase trust with observers’ perceptions of the punisher’s integrity mediating the effects of appropriate punishment on trust.*

**Punishers’ characteristics**

In addition to the circumstances of a transgression and its potential punishment, punishers’ individual characteristics can lead people to judge them as more or less trustworthy. Two obvious but particularly important characteristics are their reputation and whether their punishing acts benefit themselves.

**Reputation.** Managers’ and leaders’ beliefs and previous actions contribute to their reputations. How they have dealt with disciplinary and/or reward issues are likely to be of particular importance in this regard. We suggest that leader-punishers whose previous actions have been consistent and based on valid moral principles will have better reputations than leader-punishers whose previous actions have been inconsistent and based on pragmatism rather than valid moral principles (Butler, 1991; Butler & Cantrell, 1984). The former benefit from a predictable, value-based ethical stance; the latter suffer from being less predictable, with their
actions being attributable to a variety of possible reasons, including self-interest, revenge, or sabotage. In essence, pragmatic punishers can lose credibility when their actions are seen as convenient, capricious, and/or based on ulterior motives. Pragmatism also conflicts with deontology’s focus on duty and rights, i.e., Kant’s (1791/1952) argument that “punishment can never be administered merely as a means for promoting another good (p.397).”

Research suggests that people understand this logic. Duty, for example, often plays a more important role than consequences in shaping people’s moral judgments (DeConinck & Lewis, 1997; Hunt & Vitell, 1986) because people resist trading off their protected values (Baron & Spranca, 1997; Fiske & Tetlock, 1997). Caldwell, Pfanschmidt, and Orris (2009) also showed that employees had more trust for supervisors who took a duty-oriented approach to their ethical decisions. These findings suggest that punishers who have a principled, value-based reputation will be trusted more than punishers who have a pragmatic reputation.

Hypothesis 4 (H4): For the same punishing acts, observers will trust punishers who have a principled, value-based reputation more than punishers who have a pragmatic reputation.

Motives and self-benefit. A variety of motives can also contribute to the choice of a particular punishment. On the one hand, leaders can punish to preserve justice and the legitimacy of organizational rules and regulations. On the other, they can punish to maneuver office politics. When punishers personally benefit from punishing, observers naturally assume that self-interest motivated their behavior even if self-interest was not the only motive behind the punishment. In contrast, observers are more likely to appreciate punishment, all else equal, if it is driven by altruistic motives other than self-interest, extended kinship, or reciprocity (Fehr, Fischbacher, & Gächter, 2002; Ginitis, 2000; 2005). In particular, evolutionary theory and research have shown that going out of one’s way to punish opportunistic behavior, i.e., altruistic punishment, can
effectively foster broad systems of social cooperation because they make exploitative action costly (Boyd, et al., 2003; Fehr & Gächter, 2002; Nelissen, 2008).

Although previous research does not show that altruistic punishers engender more trust (Horita, 2010; Kiyonari & Barclay, 2008; Sterimling & Eriksson, 2014), we predict that self-benefiting punishments will diminish observers’ trust of punishers. We also suggest that, even when punishers might benefit, indirectly, from their punishment decisions, observers will still be suspicious of their motives and will be less trusting. In contrast, when self-interest is not involved, punishments should preserve trust because they signal integrity.

Hypothesis 5 (H5): For the same punishing acts, observers will trust altruistic punishers more than self-interested punishers.

The Current Research

We designed three complementary studies\(^1\) to test our hypotheses. In doing so, our research investigated the important theoretical distinctions between benevolence and integrity and how they influenced post-punishment trust in the punisher. Studies 1 and 2 used controlled scenarios, and Study 3 used a laboratory experiment to test our hypotheses. Specifically, Study 1 tested H1, H2, H3, and H5. It investigated how deservingness, punishment magnitude, and leader-punishers’ motives affected observers’ trust in punishers. Study 2 tested H4 and H5 by investigating how leader-punishers’ reputations and motives influenced people’s trust in them. Finally, Study 3 tested H5 in a controlled experiment. Using a modified Trust game (Pillutla, Malhotra, & Murnighan, 2003), it investigated punishment’s effects on trust behaviorally.

\(^1\) All three studies were reviewed and approved by the ethical review board at City University of Hong Kong (Application #: H000762, “The Paradox of Punishment and Trust”). Study 3 involved a minor deception about the existence of one player and the manipulated the risk of investing one’s own money. Participants in the first two studies were consented online but were not debriefed. Participants in Study 3 were consented in person and were generally debriefed without revealing the specifics of our study.
Study 1

Study 1 employed a 2 (deservingness: more vs. less deserving of punishment) × 3 (punishment magnitude: large vs. small vs. none) × 3 (relationship: good vs. neutral vs. bad) factorial design, with 18 different scenarios. The scenarios described a transgressor who clearly deserved either severe or light punishment. When punishments were administered, they were either large or small. Finally, to explore the effects of interpersonal biases, we also manipulated whether the punisher and the transgressor had a positive or a negative relationship (e.g., friends vs. enemies). The relationship manipulations let us test punishers’ altruistic vs. self-interested motives indirectly. Whereas under-punishing friends and over-punishing enemies both suggest self-interested motives, punishing friends and forgiving enemies are signals of integrity. We also included a neutral control condition, i.e., a normal working relationship between the two parties. We pretested the scenarios using a small student sample to make sure that they were both realistic and involving.

Participants and Procedure

We recruited 600² U.S. participants from Mturk; they averaged 36.35 years of age: 59.7% were female; 77.5% were white. Research has shown that Mturk samples tend to be a reliable population that is representative of the general public (and working people; Berinsky, Huber, & Lenz, 2012), especially when researchers prescreen the respondents (e.g., restricting participation to IP addresses located in the U.S. and using a filter to screen out people who have participated in similar studies). Mturk data also tend to be of equal or better external validity (Horton, Rand, Zeckhauser, 2011), with comparable quality, to laboratory and other online platforms (e.g.,

² We excluded 32 participants from the analyses because they failed to recall the correct information in their responses or submitted their survey from the same IP address. Including them, however, did not affect the results.
Participants read one of the 18 scenarios, randomly selected. Each scenario asked them to take the role of an observer and told them that they would need to make a decision about partnering with a punisher on a new investment project.

**Independent Variables.** In the punishment scenarios, participants were asked to play the role of an employee-observer of a leader-punisher named Pat, who was the supervisor of both the participant and another employee named James. James was described as having been Pat’s friend (good relationship), having not gotten along with Pat (bad relationship), or having a normal working relationship with Pat (neutral relationship), for many years.

During the previous two years, James either cheated on several expense reports (thereby being more deserving of punishment) or made a variety of common mistakes on his expense reports; some cost him money and some cost the company money (less deserving of punishment). As his supervisor, Pat followed company procedures and met James to discuss the issues. Following their conversation, Pat asked the company to fire James (large punishment), to give James a warning and a pay cut (medium punishment), or neither (no punishment).

All of the scenarios indicated that time had passed, and that Pat, James, and the participant had all taken jobs in other companies. But the participant had kept in contact with Pat, who recently contacted the participant about an attractive investment opportunity in an exciting new project, in a field that both Pat and the participant knew well. This promising project required strong mutual trust: if one partner was untrustworthy, there was a high likelihood that the other partner would experience a significant loss. We then asked the participants six behaviorally-oriented questions about their general trust of Pat and how likely they would be to join Pat in the investment; all of these responses were on 5-point Likert scales.
Measures

Manipulation checks. Participants answered seven manipulation check questions that assessed their memory of Pat and James’ past interaction. Two questions asked about their relationship (e.g., “how was the relationship between Pat and James?”; $\alpha=.82$); two asked about the magnitude of the punishment Pat administered (e.g., “how seriously did Pat punish James?”; $\alpha=.87$); and three asked about deservingness (e.g., “how much did James deserve to be punished?”; $\alpha=.94$).

General trust, ability, benevolence, and integrity. As noted, we used six behaviorally-oriented questions to measure participants’ general trust of Pat and the likelihood that they would make the investment ($\alpha=.97$). In addition, we adapted 16 items from Mayer and Davis (1999) to assess participants’ perceptions of Pat’s ability, benevolence, and integrity. Confirmatory factor analyses on the ratings of ability, benevolence, and integrity indicated a better model fit for the three- (CFI = .99, IFI = .99, NFI = .98, GFI = .96, RMSEA = .043) than for either of the two- (CFI = .83, IFI = .83, NFI = .82, GFI = .67, RMSEA = .17; $\Delta \chi^2 = 1171.97$, df = 2, $p < .001$) or the one-factor models (CFI = .76, IFI = .76, NFI = .75, GFI = .58, RMSEA = .20; $\Delta \chi^2 = 1670.45$, df = 3, $p < .01$). Participants’ evaluations of Pat were also reliable, for ability ($\alpha=.90$), benevolence ($\alpha=.95$) and integrity ($\alpha=.95$); we averaged each respondent’s responses to form ability, benevolence, and integrity indices.

Control variables. Because people vary in terms of their trust and punishment propensities, participants also responded to eight items on their justice sensitivity (Schmitt, Gollwitzer, & Arbach, 2005) and eight items on their trust propensity (Mayer & Davis, 1999). The justice sensitivity items ($\alpha=.82$) included questions such as “It bothers me when someone gets something s/he does not deserve” and “I am upset when someone is being treated worse than
others.” The trust propensity items (α=.71) included questions such as “One should be very cautious with strangers” and “Most people can be counted on to do what they say they will do.” All of the manipulation check, dependent, and control variable responses were on 5-point Likert scales. Finally, we asked participants to report their age, gender, marital status, education level (1=high school; 2=college; 3=graduate school), and religiousness (on scales from 1 to 10) to control for any personal differences.

**Results**

Table 1a displays the descriptive statistics and correlations for Study 1’s variables. All of the manipulations were effective. Participants rated James as being more deserving of punishment when he had cheated on his expense reports (M=4.14, SD=1.05 versus M=2.81, SD=1.01; t (565) = 15.55, p<.001, d =1.29). They also rated firing (M=4.42, SD=.80) as more consequential than a pay cut (M=3.79, SD=.81, t (373) = 7.58, p<.001, d=.78), which they rated as more consequential than no punishment (M=1.73, SD=.65, t (1, 376) = 27.31, p<.001, d=2.81). Finally, they rated the relationship between Pat and James as better when they were friends (M=4.24, SD=.69) than when they had a normal working relationship (M=3.69, SD=.74; t (369) = 7.41, p<.001, d=0.77), which they rated as better than when Pat and James did not get along (M=1.82, SD=.82; t (377) = 23.16, p<.001, d=2.39).

General trust. A 3 (punishment magnitude) × 3 (relationship) × 2 (more vs. less deserved) analysis of covariance (ANCOVA) on the general trust index, with trust propensity and justice sensitivity as covariates, led to two significant main effects - for punishment (F (2, 547) = 18.06,
and for deservingness ($F(1, 552) = 23.41, p<.001, \eta^2=.04$), but not for the two parties’ relationship ($F(2, 547) = 1.98, p=.13, \eta^2=.01$); it also led to a significant interaction between punishment and deservingness ($F(2, 547) = 48.40, p<.001, \eta^2=.15$).3

Insert Figure 1 about here

Figure 1 displays the means of the punishment-deservingness interaction. When James deserved more punishment, large ($M=3.55, SD=.97$) or medium punishment ($M=3.72, SD=.75$) led to more trust than no punishment ($M=2.38, SD=1.13$; $t(188) = 7.69; p<.001, d=1.11$; $t(183) = 9.52, p<.001, d=1.40$), with no differences in trust for large vs. medium punishment ($t(181)=1.38, p=.17, d=.20$). In contrast, when James was less deserving of punishment, large punishment ($M=2.36, SD=.98$) led to less trust than both no ($M=3.11, SD=.96$; $t(189) = 5.36, p<.001, d=-.77$) and medium punishments ($M=3.01, SD=1.03$; $t(190) = 4.51, p<.001, d=.65$), and medium and no punishment’s effects did not differ ($t(191) = .66, p=.51, d=-.10$). Thus, H1 was supported and H2 was partially supported, as medium punishments for undeserving actors did not seem to have a negative effect on people’s trust, although large punishments did.

As noted, no significant effects were found for the two parties’ relationship. Testing the effects of their relationship on trust in the two punishment conditions also yielded no significant effects (all $p’s>.50$). Thus, H5 was not supported, suggesting that people cared little about punishers’ potentially self-interested motives (i.e., under-punishing friends or over-punishing

---

3 These tests were also significant when the covariates were not included – the main effects remained significant for punishment ($F(2, 549) = 20.23, p<.001, \eta^2=.07$), and deservedness ($F(1, 549) = 22.50, p<.001, \eta^2=.04$), but not for relationship ($F(2, 549) = 2.01, p=.14, \eta^2=.01$). There were also two significant interaction effects: one between punishment and deservedness ($F(2, 549) = 49.44, p<.001, \eta^2=.15$), and one between punishment and relationship $F(4, 549) = 4.07, p=.003, \eta^2=.03$).
enemies) when they judged punishers’ trustworthiness.

**Ability, benevolence, and integrity.** General trust was highly correlated with ability ($r = .57, p < .001$), integrity ($r = .84, p < .0001$) and benevolence ($r = .72, p < .0001$). The same ANCOVA of people’s evaluations of Pat’s ability, benevolence, and integrity resulted in significant main effects for punishment ([$F (2, 552) = 9.34, p < .001, \eta^2 = .03$] for ability, [$F (2, 552) = 14.65, p < .001, \eta^2 = .05$] for benevolence, and [$F (2, 552) = 33.88, p < .001, \eta^2 = .11$] for integrity) and deservingness ([$F (1, 552) = 4.41, p = .04, \eta^2 = .02$] for ability, [$F (1, 552) = 43.03, p < .001, \eta^2 = .07$] for benevolence, and [$F (1, 552) = 29.25, p < .001, \eta^2 = .05$] for integrity), but not for their relationship.4

A significant punishment × deservingness interaction also emerged for all three dependent variables: $F (2, 552) = 6.82, p = .001, \eta^2 = .02$ for ability; $F (2, 552) = 42.76, p < .001, \eta^2 = .13$ for benevolence; and $F (2, 552) = 60.43, p < .001, \eta^2 = .18$ for integrity. In addition, there was a significant punishment × relationship interaction for integrity: $F (4, 552) = 3.33, p = .01, \eta^2 = .02$.

Figure 1 displays the means from these effects; the pattern is similar to the pattern for general trust.

**Mediation analyses.** We conducted mediation analyses to test H3 and investigate whether people’s perceptions of benevolence and integrity were the mechanisms behind these findings (see Table 1b). First, when transgressors deserved punishment, bootstrap analyses (Preacher, Rucker, & Hayes, 2007) showed that the indirect effect of punishment (effect = .65) on trust through perceived integrity was significant ($Z = 9.75, p < .001$) and zero fell outside a 95%

---

4 Again, the results were similar when the covariates were not included – the main effects remained significant for punishment ([$F (2, 549) = 10.64, p < .001, \eta^2 = .04$ for ability; $F (2, 549) = 13.12, p < .001, \eta^2 = .05$ for benevolence; and $F (2, 549) = 34.93, p < .001, \eta^2 = .11$ for integrity]), and for deservedness ([$F (1, 549) = 5.10, p = .024, \eta^2 = .01$ for ability; $F (1, 549) = 43.85, p < .001, \eta^2 = .07$ for benevolence; and $F (1, 549) = 27.88, p < .001, \eta^2 = .05$ for integrity), but not for relationship (all $p$’s > .14).
confidence interval ([.52, .77]; bias-corrected and accelerated), suggesting that perceived integrity mediated the relationship between punishment and trust. The effects were the same with people’s perceptions of ability being controlled as a covariate.

-------------------------------------

Insert Table 1b about here

-------------------------------------

When transgressors were less deserving of punishment, in contrast, bootstrap analyses revealed mediation effects for both perceived benevolence and integrity. The indirect effect of punishment (effect = -0.17) on trust through perceived benevolence was significant ($Z$=-4.42, $p<.001$) and zero fell outside a 95% confidence interval ([-.28, -.09]; bias-corrected and accelerated). The indirect effect of punishment (effect = -0.17) on trust through perceived integrity was also significant ($Z$=-3.66, $p<.001$) and zero fell outside a 95% confidence interval ([-.26, -.08], bias-corrected and accelerated. The effects remained the same with people’s perceptions of ability being controlled as a covariate.

Thus, the mediation analyses suggest that perceived integrity and benevolence mediated the relationship between punishment and trust in different situations with integrity displaying stronger effects: for more deserving transgressors, punishment increased perceptions of integrity, which in turn enhanced trust. In contrast, when transgressors were less deserving of punishment, punishment reduced perceptions of both benevolence and integrity, which damaged trust. Taken together, the results partially supported H3 as they provided full support for our prediction about the effects of integrity and limited support for our prediction about the effects of benevolence.

Taken together, these results suggest that punishing James when he deserved it increased
people’s trust of Pat (the punisher) and punishing James when he was less deserving decreased their trust of Pat. In addition, when James was more deserving of punishment, either large or medium punishment led to more trust than no punishment did. In contrast, when James was less deserving, large punishments led to less trust than either no or medium punishment. These findings provide support for H1 – that punishments in relation to deservingness increased trust – and partial support for H2, as punishing undeserving transgressors reduced trust. In addition, integrity mediated this relationship when punishment was deserved; both integrity and benevolence mediated the relationship when punishment was less deserving. These results provide partial support to H3. Pat’s potential motives behind punishment, however, did not affect trust, providing no support for H5.

**Discussion**

Study 1’s data revealed some important connections between punishment and trust. The results suggest that the effects of punishment on trust are a joint function of deservingness and punishment magnitude. Although punishing a deserving transgressor increased trust, over-punishing a less deserving transgressor decreased trust; these findings support H1 and part of H2.

The mediation analyses partially supported H3. Observers trusted leader-punishers more when transgressors were deserving because they appreciated their integrity. As a result, both moderate and large punishments enhanced trust. For less deserving transgressors, observers trusted leader-punishers less because they questioned both their benevolence and integrity. In particular, they trusted large punishers the least.

Hypothesis 5’s prediction concerning punishers’ motives was not supported, possibly because of the subtlety of our manipulation of the punisher’s motives, i.e., via their relationship with the transgressor. Thus, Study 2 used a more direct manipulation to connect the leader-
punisher’s motives with their punishing acts. Study 2 also tested H4’s prediction about the effects of the leader-punisher’s reputation. Because Study 1 showed clearly the effects of punishment magnitude, Study 2 focused only on the effects of punishment.

**Study 2**

Study 2 extended our investigation of the effects of punishment by focusing on punishers’ motives and reputations. As in Study 1, we used scenarios to test H4 and H5, i.e., whether a leader-punisher’s reputation and personal benefits from punishing affected how much observers trusted them. Specifically, H4 predicted that a principled reputation would help gain leader-punishers more trust and H5 predicted that self-benefits following a punishment would damage people’s trust in leader-punishers. To test these hypotheses, we used a 2 (punisher reputation: principled vs. flexibly pragmatic) × 2 (punishment benefiting the self vs. not) factorial design, with 4 different scenarios. Because the implicit manipulation of past relationship as a self-interested motive did not have a significant effect in Study 1, we used a more direct manipulation in this study.

**Participants and Procedure**

We recruited 160 American respondents from Mturk; they averaged 36.13 years of age: 52.38% were female; 76.87% were white. Like Study 1, each participant was randomly assigned to a scenario about partnering with a leader-punisher on a new investment project. As before, they observed a previous decision in which the leader punished an employee’s transgression.

**Manipulation of Independent Variables.** Like Study 1, we asked participants to play the role of an employee-observer of a leader-punisher named Pat, who had to deal with an employee...
who had been cheating on several expense reports over the past two years. The scenarios
described Pat as having a reputation of either being consistently value-based (e.g., by always
considering basic moral values and principles rather than the ultimate consequences) or flexible
and pragmatic (by always considering the ultimate consequences rather than core values and
basic moral principles) on disciplinary issues. Pat followed company procedures and met the
employee to discuss the issues. Pat then asked the company to fire the employee.

We also created two other conditions which described how Pat might have benefitted from
the punishing decision. In the personal benefit condition, Pat received a big promotion after the
employee was fired; in the non-benefit condition, Pat was passed over for a big promotion. Both
scenarios included explicit statements indicating that it was not clear whether Pat’s promotion or
non-promotion was related to his or her punishment recommendation.

All of the scenarios indicated that time had passed and that Pat and the participant (observer)
had moved to other jobs in other companies. The participant recently ran into Pat and they
renewed their relationship. Soon after, Pat contacted the participant about an investment project,
in a field that both Pat and the participant know well. Like Study 1, this promising project
demanded strong mutual trust: if one partner was not trustworthy, the other would have a high
likelihood of experiencing a significant loss. Also like Study 1, we asked participants two sets of
six questions each, using 5-point Likert response scales, to check our manipulations and to
measure their general trust of Pat and how likely they would join Pat in the investment.

Measures

Manipulation checks. Participants responded to 6 manipulation check questions that
assessed their memory of their past observations of Pat. Four questions asked participants about
Pat’s reputation (e.g., “Pat is flexibly pragmatic when s/he handles punishment and/or reward
issues” (1=completely disagree; 5=completely agree); two questions asked whether Pat benefitted from punishing (e.g., “How much did Pat personally benefit from the punishment recommendation?”). All of their responses were on 5-point Likert scales.

**General trust, ability, benevolence, and integrity.** Like Study 1, we used the same six questions to measure participants’ general trust of Pat and the likelihood that they would make the investment (α=.96), as well as the same 16 adapted items from Mayer and Davis (1999) to assess their perception of Pat’s ability, benevolence, and integrity. Confirmatory factor analyses on the ratings of ability, benevolence, and integrity indicated a better fit for the three- (CFI = .98, IFI = .99, NFI = .96, GFI = .94, RMSEA = .055) than for either the two- (CFI = .87, IFI = .87, NFI = .85, GFI = .74, RMSEA = .14; Δχ² = 405.46, df = 2, p < .001) or the one-factor models (CFI = .76, IFI = .76, NFI = .75, GFI = .64, RMSEA = .19; Δχ² = 796.31, df = 3, p < .001). The indices for participants’ evaluations of Pat’s ability (α=.90), benevolence (α=.92) and integrity (α=.93) were also reliable.

**Control variables.** Participants also answered Study 1’s eight items (α=.73) on their justice sensitivity (Schmitt, Gollwitzer, & Arbach, 2005) and eight items (α=.79) on their trust propensity (Mayer & Davis, 1999). Finally, we asked participants to report their age, gender, marital status, education level, and religiousness.

**Results**

Table 2a displays the descriptive statistics and correlations for the variables in Study 2. All of the manipulations were effective: people rated Pat as more principled (M=4.28, SD=.70) in the principled versus the pragmatic condition (M=3.05, SD=1.47; t (145) = 6.52, p<.001, d =1.07) and as caring more about consequences than values in the pragmatic (M=3.86, SD=.91) versus the principled condition (M=2.93, SD=.85; t (145) = 6.43, p<.001, d=1.06). They also
rated Pat as benefiting more from punishing in the self- \((M=3.43, \ SD=1.33)\) versus the non-self-benefit condition \((M=1.38,  \ SD=.83,  t (145) = 11.22,  p<.001,  d=1.85)\).

General trust. A 2 (reputation) × 2 (self-benefit) analysis of covariance (ANCOVA) on the general trust index, with trust propensity and justice sensitivity as the covariates, led to a significant main effect for self-benefit \((F (1, 141) = 18.96,  p<.001,  \eta^2=.12)\) but not for reputation \((F (1, 141) = .38,  p=.54,  \eta^2=.003)\) or their interaction, \(F (1, 143) = 2.00,  p=.16,  \eta^2=.01\).6 People trusted Pat more when the punishment involved no self-benefit than when it did \((M=3.74,  \ SD=.76 \text{ versus } M=3.07,  \ SD=1.09,  \text{respectively};  t (145) = 4.34,  p<.001,  d=.71)\). This effect was stronger for pragmatic than for principled reputations. In the pragmatic condition, Pat was seen as significantly less trustworthy after possibly benefitting from the punishment \((M=2.83,  \ SD=1.08)\) than when s/he did not benefit \((M=3.78,  \ SD=.81;  t (71) = 4.25,  p<.001,  d=-1.00)\). In the principled condition, this difference was only marginally significant \((M=3.71,  \ SD=.71 \text{ vs. } M=3.32,  \ SD=1.05;  t (72) = 1.88,  p=.06,  d=.44)\), suggesting less suspicion of Pat’s motives when s/he had a principled reputation.

Ability, Benevolence, and Integrity. The ANCOVA analyses of people’s ratings of Pat’s ability, benevolence, and integrity led to significant main effects for self-benefit on benevolence \((F (1, 143) = 15.97,  p<.001,  \eta^2=.10)\) and on integrity \((F(1, 143)=15.58, p<.001, \eta^2=.10)\); it also led to a marginally significant self-benefit × reputation interaction on benevolence, \(F (1, 143) = ________________

---

6 Like Study 1, the results were similar when the covariates were not included. The main effect was significant for self-benefit \((F (1, 143)= 19.30,  p<.0001,  \eta^2=.12)\) but not for reputation \((F(1, 143)= 1.80,  p=.18,  \eta^2=.01)\). The self-benefit × reputation interaction remained marginally significant \((F (1, 143) = 3.36,  p=.069,  \eta^2=.02)\).
Figure 2 displays these effects. In general, when the punishing decision involved no self-benefit, participants rated Pat’s benevolence ($M=3.41$, $SD=.76$ vs. $M=2.85$, $SD=.88$, $t(145)=4.19$, $p<.001$, $d=.68$) and integrity ($M=4.15$, $SD=.75$ vs. $M=3.55$, $SD=1.05$, $t(145)=3.96$, $p<.001$, $d=.66$) higher than when it did involve self-benefit. Participants’ ratings of Pat’s benevolence were also higher when Pat had a principled ($M=3.28$, $SD=.78$) rather than a pragmatic reputation ($M=2.98$, $SD=.93$, $t(145)=2.10$, $p=.037$, $d=.35$). The pattern was similar for participants’ ratings of Pat’s integrity, but the effect was only marginally significant ($M=3.99$, $SD=.86$ vs. $M=3.71$, $SD=1.04$; $t(145)=1.75$, $p=.08$, $d=.29$).

**Mediation analyses.** Like Study 1, we conducted mediation analyses to investigate whether perceptions of benevolence and integrity mediated the relationship between self-benefitting punishment and trust (see Table 2b). Bootstrap analyses (Preacher, Rucker, & Hayes, 2007) showed that the indirect effect of self-benefitting punishment (effect =-.52) on trust through perceived integrity was significant ($Z=-3.89$, $p<.001$) and zero fell outside of a 95% confidence interval ($[-.77, -.26]$; bias-corrected and accelerated). The effects were the same with perceptions of ability being controlled as a covariate. These results suggest that integrity negatively mediated the relationship between self-benefiting punishment and trust.

---

**Insert Table 2b about here**

---
Discussion

Study 2’s results suggest that a leader-punisher’s motives affect the relationship between punishment and trust. Unlike Study 1, a more direct manipulation of punishers’ self-benefits led to support for H5 in this study: when a personal benefit followed a punishment, people tended to trust leader-punishers less, even when it was not entirely clear that the benefit resulted from the punishment recommendation. Simply put, people questioned the leader’s integrity when personal benefits subsequently emerged.

In contrast, our prediction that a principled reputation would generate more trust than a pragmatic reputation was not supported. Although people tended to rate principled punishers’ benevolence and integrity higher, they did not necessarily trust them more. The marginally significant interaction effect, however, suggests that people trusted least when a pragmatic leader-punisher might have personally benefitted from punishing (see Figure 2).

Studies 1 and 2 were based on scenarios that allowed us to strictly control several contextual factors. Participants’ perceptions in these studies, however, might not be directly predictive of their behavior. Thus, Study 3 tested our hypotheses with behavioral data.

Study 3

Although Study 3 was designed to replicate the results of Studies 1 and 2 experimentally, it was markedly different in several ways. First, it gave punishers the opportunity to enact revenge for a deception that hurt them. Second, it used a behavioral measure for both punishment and trust. Third, we used a different, international sample to test the generality of these effects. Thus, we tested the effects of punishers’ motives behaviorally by manipulating and measuring self- vs. non-self-interested punishment in conditions that allowed participants to punish a potential cheater to enact revenge or to achieve public justice without self-interested revenge.
Participants and procedure

We recruited 120 participants, 65.39% female, averaging 21.55 years of age, from a major Hong Kong university. They were paid $50 HK (about $6.40 US) for their participation plus any additional payoffs that resulted from the experiment. Everyone participated in two strategic games: Game 1 was used as a punishment manipulation; Game 2 was a Trust game (see Figure 3 for the experimental procedures).

-------------------------------------
Insert Figure 3 about here
-------------------------------------

Punishment manipulations and measures (Game 1): Game 1 was based on a modified Deception Game (Erat & Gneezy, 2012; Wang & Murnighan, 2016; 2017), in which two people (Players 1 and 2) would divide HK$40 (about $5.13 US). In the game, Player 2 needed to guess a randomly generated number from 1 to 6 to determine how the two players would be paid. If Player 2’s guess was correct, Player 1 and Player 2 would each receive HK$20 (about $2.56 US); if Player 2’s guess was incorrect, Player 1 would receive HK$40 and Player 2 would get nothing.

Although both players’ payoffs depended on Player 2’s choice, Player 2s had no information about the random number. However, Player 1s knew the correct answer. Prior to Player 2’s guess, Player 1s sent a message to Player 2s in which they could reveal the correct answer or lie about it. This was the only information Player 2s had prior to their guess. The critical part of this interaction was whether Player 1s lied.

Unknown to the participants, all of them were assigned to be Player 2s in Game 1. Prior to making their guess, they received HK$10 (about US $1.3), which they could either keep for themselves or contribute to a fund that would be used to punish Player 1s if they lied. We
included two punishment conditions in the fund. In the *self-interested punishment* condition, participants were told that their HK$10 contribution would automatically punish their Player 1 counterpart (by $HK10) if s/he lied. Thus, by contributing, participants could satisfy their own, direct self-interest and take revenge on a cheating partner. In the *non-self-interested* condition, punishment was less certain and depended on how many Player 2s contributed to the fund. If every Player 2 at a session contributed, all of the Player 1s who lied would be punished. If one or more Player 2s did not contribute, however, some but not all of the Player 1s would be randomly chosen for a $10 punishment. For example, if only three out of seven Player 2s contributed and six Player 1s lied, only three of the six liars would be randomly selected to be punished. Thus, Player 2s in the non-self-interested condition could not be sure that their counterpart would be punished; direct revenge was not guaranteed as their contributions were less targeted, creating a more general punishment system.

**The Trust Game (Game 2):** After the Player 2s (i.e., all of the participants) indicated whether they would contribute $10 to the punishment fund and made their guess, they moved to Game 2 (without learning the results of Game 1). Game 2 was a Trust Game, in which all of the participants were randomly paired and randomly assigned to the role of either trustor (Player A) or trustee (Player B). In this game, trustors could invest either HK$20 or nothing of their HK$50 participation fee. If they invested nothing, the game ended. If they invested HK$20, the money would be tripled and sent to their unknown trustee/counterpart. Trustors’ choices provided a simple, dichotomous measure of trusting behavior (Fershtman, Gneezy, & List, 2012; Malhotra, 2004; Pillutla, Malhotra, & Murnighan, 2003; Snijders, 1996; Snijders & Keren, 1999). Trustees then decided whether and how much of the HK$60 to return to their unknown trustors.

Prior to making their choices, trustors (Player A) were told whether their trustee (Player B)
had contributed to Game 1’s punishment fund. This was the only information that trustors had about their trustees; it created a three-factor design: 2 (trustor vs. trustee) × 2 (Partner contributed vs. not) × 2 (self-interested vs. non-self-interested punishment system).

Prior to learning whether they would be receiving HK$20, trustees indicated how much they would return if they did receive it. Trustees who received the money then decided how much they would actually return. Although trustors were told that their final monetary outcomes depended on their choices and their counterparts’, and this was true only for people who won money. In fact, we paid anyone who lost money the full participation fee of HK$50. (No participant, however, was aware of this before the experiment finished.)

**Results**

Table 3 displays the descriptive statistics and correlations for Study 3’s variables. Most of the trustees (23 out of 28; 82%) contributed to the punishment fund in the self-interested punishment system. Far fewer (14 out of 32; 44%) contributed to the fund in the non-self-interested system ($\chi^2 = 9.31, p = .002$).

Figure 4 presents the major findings of trustors’ behavioral choices in the Trust Game. In the non-self-interested condition, 93% of the trustors (13 out of 14) sent $20 to trustees who had contributed to Game 1’s punishment fund; only 50% (9 out of 18) sent $20 to trustees who did not ($\chi^2 = 6.73, p = .009$). In the self-interested condition, in contrast, 48% of the trustors (11 out of 23) sent $20 to trustees who had contributed to the punishment fund and only 20% of the trustors (1 out of 5) sent $20 to trustees who did not ($\chi^2 = 1.30, p = .25$). Thus, punishment in the non-self-
interested punishment condition increased, as predicted. Pooling the data of trustors in the self- and non-self-interested conditions suggests that trustors trusted contributing trustees 1.5 times as often as they trusted non-contributing trustees, a marginally significant effect (24 out of 37, 65% vs. 10 out of 23, 43%; $\chi^2 = 2.64, p = .10$). These results suggest that people generally tended to trust punishers and they trusted non-self-interested punishers most.

**Additional Findings.** Controlling for whether participants contributed to the punishment fund in Game 1 in the non-self-interested condition suggests that non-contributing trustors (50%) also trusted contributing trustees more than they trusted non-contributing trustees ($\chi^2 = 5.33, p = .02$) but trustors who contributed to the fund did not trust contributing trustees more than non-contributing trustees ($\chi^2 = 1.78, p = .18$). This suggests that trustors’ preferences for punishment affected their trust of punishers; the small number of trustors in each cell, however, made this analysis only suggestive.

Although the trustees who contributed to the punishment fund were trusted more, they did not reciprocate more than the trustees who had not contributed to the punishment fund: their expected ($M_{\text{punish}} = $21.08, $SD = 15.24$ vs. $M_{\text{non-punish}} = $16.52, $SD = 13.35$, $t (58) = 1.18, p = .24, d = .32$) and real returns ($M_{\text{punish}} = $20.48, $SD = 10.71$ vs. $M_{\text{non-punish}} = $22.50, $SD = 13.35$, $t (35) = .51, p = .61, d = -.17$) did not differ significantly from each other. Thus, reciprocating trust and punishing wrongdoing may not be directly related concepts. Because we did not design this experiment to test reciprocity, however, the results are only suggestive.

**Discussion**

Study 3 replicated our previous findings using a behavioral game that allowed participants

---

7 We could not analyze the effects of trustors’ preferences in the self-interested punishment condition because there were only 4 trustors who did not contribute to the punishment fund in Game 1 and they all chose not to trust the trustee in Game 2.
to enact direct revenge. The results suggested that people who punished were trusted more than people who did not, but only when people punished in a non-self-interested manner (i.e., when revenge was not assured). In other words, people tended to trust altruistic punishers more than self-interested punishers. Thus, like Study 2, Study 3’s results supported H5, suggesting that punishers’ motives have a significant effect on people’s trust in them.

**GENERAL DISCUSSION**

This paper presents three complementary studies on the dynamic relationship between punishment and trust. The results reveal that when transgressors were deserving of punishment, either large or medium punishment was necessary for a leader-punisher to be trusted. When punishment was less deserving, however, large punishments led to less trust than either medium or no punishment, suggesting that forgiveness or leniency invited more trust in this context. A leader-punisher’s characteristics also affected people’s trust. In particular, holding punishment constant, people tended to trust altruistic punishers more than self-interested punishers. Unexpectedly, however, leader-punishers’ reputations did not have direct effects on general trust - although people thought more highly of leader-punishers’ benevolence and integrity when they had a consistently principled rather than a flexibly pragmatic reputation.

This research has important theoretical implications. First, it contributes to both the trust and punishment literatures by presenting one of the first studies of the punishment dilemma. In addition to general trust, people often judge others’ trustworthiness on three distinct dimensions: ability, benevolence, and integrity (Mayer et al., 1995; Mayer & Davis, 1999). Previous research on trust often focused on the consistency between benevolence and integrity. People who are low in integrity, for example, also tend to be low in benevolence. Thus, to be trustworthy, a leader must be attributed with both high integrity and benevolence. In organizational life, however,
benevolence and integrity do not always coincide, e.g., when ethical actions like punishment result in non-benevolent outcomes. Our research identifies the mechanisms behind people’s trust and distrust when they face these two conflicting moral values. For observers, its resolution sends strong signals about leader-punishers’ intentions. While punishing transgressions is consistent with integrity, its inevitable harm can also create less than benevolent outcomes. Addressing transgressions by confronting the punishment dilemma is particularly important in organizations because managers cannot be consistently effective if inappropriate and immoral behavior goes unpunished or if employees feel that they are cruel and harsh and lack sympathy and humanity.

By focusing on the important social effects of trust (Treviño, 1992), our investigation suggests that people sometimes find it difficult to separate benevolence and integrity (Colquitt et al., 2012). Instead, they may exhibit a halo effect, with highly correlated perceptions of trust, benevolence, and integrity. At the same time, the current studies show that people can distinguish between benevolence and integrity - when transgressors are more deserving of punishment. When punishment was less deserving, however, the effects of integrity and benevolence tended to commingle. This suggests that people do not consistently distinguish these two trusting attributes when they intuitively judge the trustworthiness of leader-punishers. Instead, their perceptions of benevolence and integrity seem more differentially sensitive when punishment is warranted rather than when it is not.

Second, our research is among the first set of empirical studies to investigate the dynamic relationship between punishment and trust. Prior organizational research has investigated different (and important) effects of punishments and rewards (Atwater, et al., 1997; Podsakoff et al., 2006). By focusing on the important contextual conditions of punishment and inherent characteristics of leader-punishers, the current research tests how punishment interacts with these factors to
influence people’s trust of a leader-punisher. Like the predictions of just deserts theory, our results suggest that the effects of punishment on trust are sensitive to punishers’ characteristics and motives as well as contextual factors such as deservingness and the magnitude of punishment. These findings provide important insights into the dynamic relationship between punishment and trust in organizations. Future research might investigate some of the additional dynamics of these effects, especially their relationship with other important variables such as organizational cultures and norms.

**Practical Implications**

The current research also provides important practical implications concerning how managers and leaders can administer punishment without jeopardizing trust in themselves. Punishment is a highly sensitive issue in organizational life. Managers and leaders often use punishment but they may be reluctant to admit it (Sim, 1980) because punishing suggests non-benevolence. The current findings on the interaction effects between deservingness and punishment suggest that organizations should carefully consider how, when, and why they administer punishment. In particular, organizations might want to establish clear standards for both deservingness and the magnitude of punishment (e.g., by providing detailed criteria and past examples). They may also consider transparently publicizing these standards in organizational policy or codes of conduct. In doing so, managers can then rely on objective criteria instead of their intuitive judgments when they decide whether transgressors are deserving of punishment. When they conclude that punishment should be imposed, they should reserve large, harsh punishment only for truly deserving transgressors, as this will not only make other employees view the process as being fair (Treviño, 1992; Treviño & Ball, 1992); they are also likely to increase their trust in their leaders’ integrity. For less deserving transgressors, however, managers might be advised to show more
forgiveness and lenience. In particular, they should be warned against using large punishments which can damage other employees’ trust in their benevolence. Instead, punishment for less deserving wrongdoers should tend to be only small or moderate.

When managers or leaders lack clear guidelines or standards to judge transgressors’ deservingness, our research suggests that moderate punishment may often be preferred over serious punishments especially when employees’ trust is a critical organizational issue. Our results also suggest that, for deserving targets, punishment increased trust more than non-punishment did, but punishment severity (large vs medium in the current studies) did not affect observers’ trust perceptions. These effects extended to less deserving targets, where large punishments damaged trust but medium punishments did not. Thus, although our research only provides initial, suggestive findings, it does suggest that managers or leaders may be well advised to consider using moderate punishment when they are uncertain about how deserving transgressors are for punishment.

Second, our results suggest that organizations might take care in appointing personnel to administer punishment, especially when self-benefit is potentially involved. If, for example, managers can benefit by punishing, even indirectly or unintentionally, they probably should not be asked to administer or recommend the punishment. Managers who are known to be flexibly pragmatic are also likely to be poor choices as punishers, as other people are likely to question both their integrity and their benevolence. Thus, organizations might sometimes consider using a rotating team of independent managers as a ‘punishment panel. Such a panel could apply appropriate punishments without implying the presence of ulterior motives.

Finally, our research suggests that organizations may need to consider the overall social impact of punishing transgressors, with a special emphasis on how impaired trust can be restored.
Because punishment is a double-edged sword with affective, attitudinal, and behavioral consequences (Arvey & Ivancevich, 1980; Treviño, 1992), it can potentially damage overall trust in an organization. When punishing wrongdoing damages trust, it is important to repair and restore trust to rebuild employees’ confidence in the organization’s disciplinary system. In particular, our research suggests that benevolence and integrity are two essential elements that organizations may need to promote when they seek to restore impaired trust. For example, when employees question the integrity of leader-punishers, organizations may need to take necessary actions to reestablish their integrity, possibly via avenues unrelated to discipline, as integrity depends on many factors. When employees are suspicious of their leader-punishers’ benevolence, in contrast, organizations might consider whether their disciplinary system is callous and/or inappropriate. Research is necessary to determine whether these kinds of adjustments are effective. At a minimum, however, they are likely to provide useful strategies for aligning punishments with an organization’s attempts to improve employees’ trust.

Limitations and Future Directions

Like all studies, the current research has limitations. Because previous research showed that punishment can have cognitive, affective and behavioral impacts on general observers (e.g., Niehoff et al., 1998; O’Reilly & Puffer, 1989; Treviño, 1992; Treviño & Ball, 1992), our investigation also focused on observers who had no direct relationship with either punishers or transgressors. In real organizations, however, observers can have strong relationships with either or both parties. The typical example is punishment within small teams. For example, after a team leader punishes a team member, the leader and the punished member will still likely have to
work closely with each other in the same team\(^8\). In this context, the leader’s punishments can influence both leader-member relations and team culture. Our current research does not extend to overall trust within a team; this is an important topic worthy of future investigation.

Team-oriented research on punishment and trust would also provide an opportunity to bridge interdisciplinary theories on team dynamics and punishment. In particular, in addition to the punisher’s characteristics and the deservingness of the punished person, the punished person’s post-punishment reactions toward both the punisher and other teammates are likely to reflect the effectiveness of punishment, which may in turn affect other teammates’ trust in the punisher-leader. Reintegrative reform theories of punishment suggest that punishment can serve penitential and educational functions that help punished people return to their communities and groups (Duff, 1986; Hampton, 1984; Reitan, 1996; Treviño and Weaver, 1998). A punished person’s penance and reparation, for example, can not only facilitate this reintegration process; it might also enhance other team members’ trust in their leader. In contrast, negative reactions like animosity, malice, revenge, and sabotage might signal that the punishment was not effective, thereby stimulating team members’ distrust of their leader’s competence.

Thus, the interactional aspects of punishment in a team, e.g., reintegration, are likely to create additional trust dilemmas that pit integrity not only against benevolence, but also against ability, potentially as a function of the punished wrongdoer’s reactions. Previous research has not adequately investigated the reintegration of punished people into their teams, groups, and organizations. Treviño and Weaver (1998), for instance, have noted that organizational research would benefit from incorporating reintegrative theories of punishment. We second their call and

\(^8\) We thank an anonymous reviewer for identifying this important issue.
suggest that it might be particularly worthwhile to investigate how transgressors react to their punishment and how the process of reintegration influences the relationship between punishment and trust in groups and teams.

Obviously, other organizational factors can also influence employees’ trust toward their leader-punishers. People who have a close relationship with a punished person inside or outside of work are likely to exhibit enhanced sympathy toward the punished person and, as a result, less trust toward the punisher. Similarly, people who have conflicts of interest with a punished person, e.g., direct supervisors or subordinates, are also likely to display biased trust toward either the punisher or the punished person when their interests are affected. When organizational politics are involved, punishing one transgressor may hurt an entire work group. The current research could not address these broader issues. Our hope is that future research will investigate additional factors to broaden our understanding of these central team and organizational dynamics.

Methodologically, our first two studies used controlled scenarios that measured self-reported perceptions rather than behavior. In addition, our manipulations of punishers’ motives may have had particular contextual effects on people’s intuitive, trusting judgments. Also, we only used a male name, James, for the punished person. Although we chose a sex-neutral name, Pat, for the punisher, we did not measure how people interpreted the punisher’s sex. Research on gender stereotypes (e.g., Heilman et al., 2004; Heilman, & Okimoto, 2007) suggests that female punishers may be trusted less because punishment does not fit a typical female stereotype. Thus, the question of whether sex might influence people’s trust in a punisher is another question.

---

9 We thank an anonymous reviewer for identifying this new research question.
deserving future research. Finally, the lack of significant effects of leader-punishers’ reputations on trust suggests, intuitively, that further work is necessary on this issue, as past research has shown that reputation has a positive effect in stimulating fairness and cooperation (Sigmund, Hauert, & Nowak, 2001).

Although Study 3’s behavioral measure of trust led to findings that were consistent with those of the first two studies, it was difficult to experimentally manipulate all of the factors that we included in Studies 1 and 2. Also, letting people make several choices reduced the sample size in some conditions, limiting the strength of the findings. In addition, the specifics of the Trust Game might also limit its immediate generalizability to richer organizational contexts (Murnighan & Wang, 2016).

**Conclusions**

In spite of these limitations, three complementary studies found consistent results on the important, dynamic relationship between punishment and trust. Punishing deserving perpetrators increased people’s trust but over-punishing less deserving perpetrators had the opposite effect. Punishment magnitude also had a considerable influence on people’s trust as both large and medium punishment for deserving perpetrators enhanced trust but only large punishment for less deserving perpetrators damaged it (medium punishment did not). Punishers’ motives played an additional, important role in influencing trust, with people trusting altruistic leader-punishers more than potentially self-interested punishers. Taken together, these findings indicate that punishment can enhance or damage trust depending on critical contextual factors that effective organizations will want to understand.
References


Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?:


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


Kant, I. (1952). The science of right (W. Hastie, Trans.). (Original work published 1797) In R.
Hutchins (Ed.), *Great books of the Western World* (Vol. 42, pp. 397-446). Chicago, IL:
Encyclopedia Brittanica.


The effects of punishment on people’s average ratings of general trust, ability, benevolence, and integrity (Study 1\textsuperscript{10}).

\textsuperscript{10} We only included the conditions of punishment magnitude and deservingness as we did not find any significant effects for the two parties’ relationship.
Figure 2.

The effects of punishment and reputation on people’s average ratings of general trust, ability, benevolence, and integrity (Study 2).
Figure 3.

Procedure of Experiment 3.

In Game 1, participants chose whether they would contribute to a punishment fund, which differed in the two experimental conditions. In Game 2, trustors decided how much they would trust trustees after learning whether their trustees had contributed to the punishment fund in Game 1.
Figure 4.
Percentages of trustors trusting trustees in the second game, the Trust Game, as a function of the conditions and whether trustees had previously contributed to the punishment fund for the first game, the Deception Game (Study 3).
Table 1a

Descriptive Statistics and Correlations Among Studied Variables (Study 1)

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gender</td>
<td>1.60</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Age</td>
<td>36.35</td>
<td>12.88</td>
<td>.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Marital status</td>
<td>1.59</td>
<td>.49</td>
<td>-.04</td>
<td>-.23**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Education</td>
<td>1.89</td>
<td>.60</td>
<td>.01</td>
<td>.05</td>
<td>-.13**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Religiousness</td>
<td>4.66</td>
<td>3.37</td>
<td>.09*</td>
<td>.20**</td>
<td>-.22**</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>6. General trust</td>
<td>3.01</td>
<td>1.11</td>
<td>-.10*</td>
<td>-.00</td>
<td>-.04</td>
<td>-.08</td>
<td>-.05</td>
</tr>
<tr>
<td>7. Ability</td>
<td>3.59</td>
<td>.71</td>
<td>.05</td>
<td>-.02</td>
<td>.01</td>
<td>-.11**</td>
<td>.01</td>
</tr>
<tr>
<td>8. Benevolence</td>
<td>3.08</td>
<td>.94</td>
<td>-.06</td>
<td>.01</td>
<td>-.05</td>
<td>-.12**</td>
<td>.02</td>
</tr>
<tr>
<td>9. Integrity</td>
<td>3.19</td>
<td>1.12</td>
<td>-.08</td>
<td>.01</td>
<td>-.05</td>
<td>-.06</td>
<td>-.04</td>
</tr>
<tr>
<td>10. Deservingness</td>
<td>.49</td>
<td>.50</td>
<td>-.02</td>
<td>.02</td>
<td>-.03</td>
<td>-.02</td>
<td>-.03</td>
</tr>
<tr>
<td>11. Relationship</td>
<td>1.99</td>
<td>.82</td>
<td>-.05</td>
<td>-.01</td>
<td>.00</td>
<td>-.10*</td>
<td>-.01</td>
</tr>
<tr>
<td>12. Trust propensity</td>
<td>2.73</td>
<td>.49</td>
<td>-.07</td>
<td>.12**</td>
<td>-.01</td>
<td>.04</td>
<td>-.01</td>
</tr>
<tr>
<td>13. Justice sensitivity</td>
<td>3.79</td>
<td>.58</td>
<td>.21**</td>
<td>-.05</td>
<td>.03</td>
<td>-.03</td>
<td>-.01</td>
</tr>
</tbody>
</table>

Note. N=568, **p<.01; *p<.05
Table 1a (Continued)

*Descriptive Statistics and Correlations Among Studied Variables (Study 1)*

<table>
<thead>
<tr>
<th>Variable</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. General trust</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Ability</td>
<td>.57**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Benevolence</td>
<td>.72**</td>
<td>.46**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Integrity</td>
<td>.84**</td>
<td>.56**</td>
<td>.71**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Deservingness</td>
<td>.17**</td>
<td>.09*</td>
<td>.24**</td>
<td>.18**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Relationship</td>
<td>-.09</td>
<td>-.05</td>
<td>-.05</td>
<td>-.09*</td>
<td>-.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Trust Propensity</td>
<td>.23**</td>
<td>.17**</td>
<td>.19**</td>
<td>.21**</td>
<td>.02</td>
<td>-.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Justice sensitivity</td>
<td>-.09</td>
<td>-.03</td>
<td>-.07</td>
<td>-.04</td>
<td>-.02</td>
<td>-.01</td>
<td>-.19**</td>
<td></td>
</tr>
</tbody>
</table>

*Note. N=568, **p<.01; *p<.05*
Table 1b

Results of Mediation Analysis (Study 1)

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>SE</th>
<th>Bias corrected 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td><strong>When transgressors deserved punishment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punishment to integrity (a path)</td>
<td>.77***</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>Integrity to trust (b path)</td>
<td>.85***</td>
<td>.04</td>
<td></td>
</tr>
<tr>
<td>Punishment on trust (c path)</td>
<td>.59***</td>
<td>.08</td>
<td></td>
</tr>
<tr>
<td>Punishment on trust (c’ path)</td>
<td>-.06</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Indirect effect through integrity (ab path)</td>
<td>.65***</td>
<td>.07</td>
<td>.52</td>
</tr>
<tr>
<td><strong>When transgressors were less deserving of punishment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punishment to benevolence (a path)</td>
<td>-.57***</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>Punishment to integrity (a path)</td>
<td>-.28***</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>Benevolence to trust (b path)</td>
<td>.30***</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>Integrity to trust (b path)</td>
<td>.60***</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Punishment on trust (c path)</td>
<td>-.38***</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>Punishment on trust (c’ path)</td>
<td>-.04</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Indirect effects through benevolence (ab path)</td>
<td>-.17***</td>
<td>.04</td>
<td>-.27</td>
</tr>
<tr>
<td>Indirect effects through integrity (ab path)</td>
<td>-.17***</td>
<td>.05</td>
<td>-.27</td>
</tr>
</tbody>
</table>

Note. Mediation effects were estimated by bootstrap analyses (5000 bootstraps).

***p<.001; **p<.01; *p<.05
Table 2a

Descriptive Statistics and Correlations Among Studied Variables (Study 2)

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gender</td>
<td>1.52</td>
<td>.50</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Age</td>
<td>36.13</td>
<td>12.06</td>
<td>.21*</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Marital status</td>
<td>1.66</td>
<td>.48</td>
<td>-.11</td>
<td>-.25**</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Education</td>
<td>1.95</td>
<td>.67</td>
<td>-.01</td>
<td>.07</td>
<td>-.27**</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>5. Religiousness</td>
<td>4.44</td>
<td>3.45</td>
<td>.33**</td>
<td>.16</td>
<td>-.21*</td>
<td>.06</td>
<td>–</td>
</tr>
<tr>
<td>6. General trust</td>
<td>3.41</td>
<td>.99</td>
<td>-.05</td>
<td>.03</td>
<td>-.06</td>
<td>-.06</td>
<td>-.00</td>
</tr>
<tr>
<td>7. Ability</td>
<td>3.83</td>
<td>.67</td>
<td>.08</td>
<td>-.01</td>
<td>-.02</td>
<td>-.23**</td>
<td>-.06</td>
</tr>
<tr>
<td>8. Benevolence</td>
<td>3.13</td>
<td>.87</td>
<td>-.02</td>
<td>-.02</td>
<td>-.09</td>
<td>-.11</td>
<td>-.01</td>
</tr>
<tr>
<td>9. Integrity</td>
<td>3.85</td>
<td>.96</td>
<td>.11</td>
<td>.01</td>
<td>.00</td>
<td>-.05</td>
<td>.02</td>
</tr>
<tr>
<td>10. Reputation</td>
<td>.50</td>
<td>.50</td>
<td>-.10</td>
<td>.08</td>
<td>.12</td>
<td>.07</td>
<td>-.11</td>
</tr>
<tr>
<td>11. Self-benefit</td>
<td>.50</td>
<td>.50</td>
<td>.01</td>
<td>.19*</td>
<td>-.02</td>
<td>.03</td>
<td>.04</td>
</tr>
<tr>
<td>12. Trust propensity</td>
<td>2.35</td>
<td>.48</td>
<td>-.08</td>
<td>.07</td>
<td>-.07</td>
<td>.01</td>
<td>-.18*</td>
</tr>
<tr>
<td>13. Justice sensitivity</td>
<td>3.85</td>
<td>.56</td>
<td>.26**</td>
<td>.12</td>
<td>-.04</td>
<td>-.07</td>
<td>.06</td>
</tr>
</tbody>
</table>

Note. N=147, **p<.01; *p<.05
Table 2a (Continued)

Descriptive Statistics and Correlations Among Studied Variables (Study 2)

<table>
<thead>
<tr>
<th>Variable</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. General trust</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Ability</td>
<td>.59**</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Benevolence</td>
<td>.71**</td>
<td>.43**</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Integrity</td>
<td>.86**</td>
<td>.52**</td>
<td>.75**</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Reputation</td>
<td>.11</td>
<td>-.07</td>
<td>.17*</td>
<td>.14</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Self-benefit</td>
<td>-.34**</td>
<td>-.08</td>
<td>-.33**</td>
<td>-.31**</td>
<td>-.01</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Trust Propensity</td>
<td>.23**</td>
<td>.02</td>
<td>.36**</td>
<td>.23**</td>
<td>.27**</td>
<td>-.12</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>13. Justice sensitivity</td>
<td>.17*</td>
<td>.21*</td>
<td>-.02</td>
<td>.16</td>
<td>.00</td>
<td>.06</td>
<td>-.17*</td>
<td>–</td>
</tr>
</tbody>
</table>

Note. N=147, **p<.01; *p<.05
### Table 2b

*Results of Mediation Analysis (Study 2)*

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>SE</th>
<th>Bias corrected 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>Self-benefiting punishment to</td>
<td>-.60***</td>
<td>.15</td>
<td></td>
</tr>
<tr>
<td>integrity <em>(a path)</em></td>
<td></td>
<td></td>
<td>Upper</td>
</tr>
<tr>
<td>Integrity to trust <em>(b path)</em></td>
<td>.87***</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Self-benefiting punishment</td>
<td>-.67***</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td>on trust <em>(c path)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-benefiting punishment</td>
<td>-.16†</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>on trust <em>(c’ path)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirect effects through</td>
<td>-.52***</td>
<td>.04</td>
<td>-.77</td>
</tr>
<tr>
<td>integrity <em>(ab path)</em></td>
<td></td>
<td></td>
<td>-.27</td>
</tr>
</tbody>
</table>

*Note. Mediation effects were estimated by bootstrap analyses (5000 bootstraps).*
Table 3

Descriptive Statistics and Correlations Among Studied Variables (Study 3)

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Age</td>
<td>21.80</td>
<td>3.02</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Gender</td>
<td>1.71</td>
<td>.46</td>
<td>-.07</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Trust</td>
<td>12.00</td>
<td>9.87</td>
<td>-.03</td>
<td>-.04</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Partner contributed</td>
<td>.62</td>
<td>.49</td>
<td>.03</td>
<td>.05</td>
<td>.21</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>5. Punishment system</td>
<td>1.54</td>
<td>.50</td>
<td>-.01</td>
<td>-.12</td>
<td>.25*</td>
<td>-.39**</td>
<td>–</td>
</tr>
</tbody>
</table>

Note. N=60, **p<.01; *p<.05