On Sanction-Goal Justifications: How and Why Deterrence Justifications Undermine Rule Compliance

Marlon Mooijman
University of Southern California

Wilco W. van Dijk and Eric van Dijk
Leiden University

Naomi Ellemers
Utrecht University

Authorities frequently justify their sanctions as attempts to deter people from rule breaking. Although providing a sanction justification seems appealing and harmless, we propose that a deterrence justification decreases the extent to which sanctions are effective in promoting rule compliance. We develop a theoretical model that specifies how and why this occurs. Consistent with our model, 5 experiments demonstrated that—compared with sanctions provided without a justification or sanctions provided with a just-deserts justification—sanction effectiveness decreased when sanctions were justified as attempts to deter people from rule breaking. This effect was mediated by people feeling distrusted by the authority. We further demonstrated that (a) the degree to which deterrence fostered distrust was attenuated when the sanction was targeted at others (instead of the participant) and (b) the degree to which distrust undermined rule compliance was attenuated when the authority was perceived as legitimate. We discuss the practical implications for authorities tasked with promoting rule compliance, and the theoretical implications for the literature on sanctions, distrust, and rule compliance.

Keywords: deterrence, distrust, rule compliance, sanction justifications

Authorities often provide a justification for their sanctioning behavior. Judges sentence people to prison with the explicit justification that this is meant to deter future rule-breaking behavior (e.g., see Martinez, 2015) and politicians explicitly justify naming-and-shaming policies as attempts to deter crime (e.g., see Langlois, 2012). In the present research, we investigate the consequences of providing such deterrence justifications. We propose that deterrence justifications decrease the extent to which sanctions are effective in promoting rule compliance and we propose that this can be attributed to people feeling distrusted by authorities that justify their sanctions as deterrents.

Previous research has mainly focused on the extent to which sanction goals such as deterrence guide sanctioning decisions (Carlsmith, 2006, 2008; Carlsmith, Darley, & Robinson, 2002; Darley, Carlsmith, & Robinson, 2000; Gerber & Jackson, 2013), but has left the effects of sanction-goal justifications on rule compliance unaddressed. Examining how and why deterrence justifications shape people's willingness to comply with rules can provide valuable insights into how authorities should—and should not—use sanctions. Indeed, societal and organizational authorities (e.g., policymakers, leaders, and managers) tend to justify their use of sanctions by stressing the necessity to deter rule-breaking behavior (Kirchler, Kogler, & Muehlbacher, 2014; Mooijman, van Dijk, Ellemers, & van Dijk, 2015). Understanding how such justifications affect rule compliance may therefore be helpful in explaining the (in)effectiveness of real-life sanctions and suggesting ways to improve the manner in which authorities justify their sanctioning behavior.

Sanction Justifications

Authorities frequently stress the aim to deter rule breaking as justification for their use of sanctions (Bentham, 1789/1988; Hobbes, 1651/1988; Kirchler et al., 2014; Mooijman et al., 2015; Nagin, 1998). When having this sanction goal, authorities should be primarily concerned with using sanctions to deter future rule breaking from potential rule breakers rather than punishing past rule breakers proportionate to their crime (Carlsmith et al., 2002). A deterrence goal is thus prospective rather than retroactive and can, as such, be distinguished from just deserts—a retroactive sanction goal (Darley, 2009; Kant, 1780/1961). When having just deserts as the sanction goal, authorities should be primarily concerned with punishing rule breakers proportionately for crimes committed in the past (i.e., achieve balance between crime and

This article was published Online First December 1, 2016.

Marlon Mooijman, Department of Psychology, University of Southern California; Wilco W. van Dijk and Eric van Dijk, Department of Social and Organizational Psychology, Leiden University; Naomi Ellemers, Department of Social, Health, & Organizational Psychology, Utrecht University.

Correspondence concerning this article should be addressed to Marlon Mooijman, Department of Psychology, University of Southern California, SGM 501, 3620 South McClintock Avenue, Los Angeles, CA 90089-1061. E-mail: mooijman@usc.edu

http://dx.doi.org/10.1037/pspi0000084

punishment), regardless of the sanction's ability to deter future rule breaking (Carlsmith et al., 2002; Keller, Oswald, Stucki, & Gollwitzer, 2010).

Although reliance on deterrence as a sanction goal can affect the type and severity of the sanction used (Mooijman et al., 2015; Tetlock et al., 2007) and thereby influence rule compliance (Ball, Treviño, & Sims, 1994), we argue that authorities' use of a deterrence goal as *justification* creates an additional source of influence. That is, independently of the type and severity of a sanction, people's willingness to comply with rules may be negatively affected by whether an authority provides a deterrence justification or not. We propose that this negative impact of deterrence justifications is specific to deterrence justifications and does not similarly hold for just-deserts justifications. Understanding the specificity of the effect is important because scholars typically contrast deterrence with just deserts sanction goals and thus study these two sanction goals simultaneously (Bentham, 1789/1988; Carlsmith et al., 2002; Hobbes, 1651/1988; Kant, 1780/1961; Keller et al., 2010; Nagin, 1998).

When Sanction Justifications May Signal Distrust

A central aspect of a deterrence goal is that sanctions are aimed at those who are deemed likely to break rules (hence the need to deter them; Nagin, 1998). In other words, authorities that aim to deter rule breaking are focused on the possibility that people break rules in the future (i.e., they distrust them; Mooijman et al., 2015). In contrast, the goal to give past rule breakers their just deserts is indifferent with regard to people's likelihood of breaking rules in the future (i.e., trustworthiness is irrelevant; Carlsmith et al., 2002; Kant, 1780/1961; Mooijman et al., 2015).

People often infer intentions and considerations from authorities' decisions McKenzie, Liersch, and Finkelstein (2006). Managers' attempts to incentivize weight-loss with financial sanctions have been shown, for instance, to unintentionally signal negative attitudes toward the overweight (Tannenbaum, Valasek, Knowles, & Ditto, 2013). Authorities who justify sanctions attempts to deter people from rule breaking may therefore signal their distrust to people. That is, authorities using deterrence justifications may signal that sanctions are needed because people are likely to break rules in the absence of sanctions; sanctions are then used as a means to deter people's future rule-breaking behavior. The communicated "breadth" of deterrence-justified sanctions is thus large (i.e., targets all potential rule breakers), thereby signaling distrust to those who have not broken any guidelines or rules (yet). This distrust-signaling effect of sanction justifications may be specific to deterrence. Just deserts signals that authorities' sanctions are aimed only at those who have broken rules in the past instead of those who may potentially break rules in the future. This reduces the likelihood that people who have not broken any rules yet feel distrusted.

Evidence corroborating this reasoning comes from research showing that authorities' distrust predicts the degree to which they rely on deterrence, but not just deserts, as a sanction goal (Mooijman et al., 2015) and from research showing that people are highly motivated to infer authorities' considerations from their sanction decisions (Fiske, 1993; Keltner et al., 2003). People may thus infer from authorities' deterrence justifications that they are expected to have the malicious intention to undermine the interests of the authorities (consistent with definitions of distrust, Kramer, 1999; Yamagishi & Yamagishi, 1994; Zand, 1997). In sum, we hypothesize that justifying a sanction as an attempt to deter people from breaking rules increases the degree to which people feel distrusted by the relevant authorities (Hypothesis 1).

Why Feeling Distrusted May Undermine Rule Compliance

How is people's rule compliance affected by their feelings of being distrusted by the authorities? Rule compliance is not solely determined by the severity of a sanction or the probability that one receives a sanction (Balliet, Mulder, & Van Lange, 2011). Instead, rule compliance is also determined by how people feel treated by authorities (e.g., interpersonal justice; Tyler & Lind, 1992). For instance, people's satisfaction with authorities' decisions decreases when authorities communicate disrespect through pursuing their own interest instead of the interest of the people (De Cremer & Van Knippenberg, 2002) and using nontransparent and biased procedures (Tyler, 2006). In contrast, authorities that are perceived to pursue the collective interest (Mulder & Nelissen, 2010), show respect for others (Tyler & Blader, 2003), and use transparent and unbiased procedures foster decision acceptance (Tyler, 2006). Importantly, these effects of perceived interpersonal treatment often go beyond the outcome that people (expect to) receive from authorities (Cropanzano et al., 2007). How people feel treated by authorities justifying sanctions may thus be of vital importance for people's willingness to behave according to authorities' rules.

More specifically, people are motivated to see themselves as trustworthy (Brown, 2012; Sedikides, Meek, Alicke, & Taylor, 2014; Steele, 1988), and want and expect others to trust them (Ellemers, 2012; Tyler & Lind, 1992). Feeling *dis*trusted by an authority is therefore likely to foster the feeling that the authority does not view oneself favorably (e.g., without respect). This perception alone may be sufficient to feel poorly treated, thereby undermining one's willingness to abide to this authority's rules. Indeed, perceived interpersonal treatment does not have to revolve around tangible outcomes that one receives, but can also entail subjective assessments of how others view oneself (e.g., respect; Ellemers, Doosje, & Spears, 2004). A perceived lack of trust in one's willingness to comply with relevant guidelines and rules may seem unwarranted when no prior breach of rules was displayed, and may thus seem disrespectful and unjust. This may in turn decrease people's willingness to comply with rules—research has indeed demonstrated that even slight signs of distrust (i.e., when senders in a Trust Game do not sent their full endowment) can increase interpersonal retaliation in Trust Games (Pillutla, Malhotra, & Murnighan, 2003). Although sanctions increase the costs of rule breaking regardless of an authority's sanction justification, we thus predict that the potential effectiveness of a sanction is decreased by the distrust that people experience when an authority provides a deterrence justification.

This prediction can also be construed as a behavioral confirmation effect (i.e., self-fulfilling prophecy)—people's willingness to comply with authorities' rules is expected to decrease in response to people's perception that these authorities expect them to break rules. Research has demonstrated that authorities can create self-fulfilling prophecies; subordinates are more likely to be productive when supervisors expect them to be intrinsically motivated and

productive, in part because supervisors' set behavioral standards through their expectations (Pelletier & Vallerand, 1996). Although our theorizing on *how* authorities create behavioral confirmation effects is different from this previous research (i.e., we focus on deterrence justifications and the role of feeling distrusted), our predictions can be construed as authorities creating a self-fulfilling prophecy.

According to Snyder (1992) a self-fulfilling prophecy has four stages: (a) The perceiver adopts certain beliefs about the target, (b) the perceiver behaves as if these beliefs were true and treats the target accordingly, (c) the target perceives and responds to the perceiver's behaviors, and (d) the perceiver interprets the target's behaviors as a confirmation of his or her initial beliefs. Previous research on power and sanctions fits into the first two stages of the model—(a) the power that authorities hold has been shown to increase distrust (Inesi, Gruenfeld, & Galinsky, 2012; Mooijman et al., 2015; Schilke, Reimann, & Cook, 2015), and (b) authorities justify their sanctions as means to deter rule breaking because of this distrust (Mooijman et al., 2015). The theorizing presented in the current manuscript fits primarily in the third stage (c)—people infer distrust from authorities' deterrence justifications and are consequently less willing to comply with authorities' rules. Although the current research does not directly test the fourth and final stage (the perceiver interprets the target's behaviors as a confirmation of his or her initial beliefs), it seems likely that people's unwillingness to comply with rules confirms authorities' initial distrust.

In sum, we hypothesize that sanction effectiveness decreases when authorities justify their sanctions as a means to deter people from breaking rules (Hypothesis 2a). We further hypothesize that distrust mediates this negative relationship between deterrence justifications and rule compliance (Hypothesis 2b).

Overview of Current Research

We tested our hypotheses in five experiments in which we, (a) manipulated whether or not an authority provided a deterrence justification for its sanctioning behavior, (b) measured the distrust that participants felt, and (c) tested how this affected rule compliance. In most experiments, we thus compared a condition in which an authority justified a sanction as a means to deter to a condition in which an authority provided no justification for a sanction. In two of the experiments, we included a condition in which authorities provided a just-deserts justification or provided both a deterrence *and* just deserts justification. This allowed us to demonstrate that the theorized effects of sanction justifications are specific to deterrence justifications. Moreover, in these experiments we tested how deterrence justification affected the extent to which participants lied to their team leader to further their own interest (Experiments 1 and 2), their willingness to commit plagiarism (Experiment 3) or fraud (Experiment 4), and their willingness to take resources from their leader (Experiment 5). We tested our hypotheses in both college samples (Experiments 1 and 4) and Mechanical Turk samples (Experiments 2, 3, and 5).

To provide support for the proposed mediating role of distrust and exclude alternative explanations, we assessed the perceived anger of the authority, attitudes toward the authority, and participants' distrust toward others. Previous research has demonstrated that rule compliance can be affected by the extent to which people perceive others to be angry (Wubben, De Cremer, & van Dijk, 2011), the attitudes that they hold toward authorities (Tyler & Blader, 2003) and the extent to which they trust others to comply with rules (Mulder, van Dijk, De Cremer, & Wilke, 2006). We assessed these control variables in Experiments 2, 3, and 4. In Experiments 4 and 5, we examined the moderating role of authority legitimacy and perceived target of the sanction (i.e., oneself or others). Based on the notion that deterrence justifications increase distrust because of their "breadth" (i.e., they cast a wide net of suspicion that includes the participant), we expect the relationship between deterrence and distrust to be attenuated when the sanction is perceived to target others instead of oneself. Moreover, based on the notion that distrust undermines rule compliance because of the way people feel treated by the authority, we expect legitimacy to attenuate the degree to which distrust undermines rule compliance; legitimacy has been shown to buffer against relational threats (Tyler, 2006).

Consistent with the recommendations of Simmons, Nelson, and Simonsohn (2011), we made sure that every condition had around 30 participants, although most conducted experiments have considerably more participants per condition (i.e., more than 50; cf. Simmons, Nelson, & Simonsohn, 2013). Across experiments, participants reported very low rates of suspicion regarding the goal of the experimental manipulations (<5%); results did not change significantly for any experiment when we excluded participants who were suspicious. Consequently, we report the analyses that include the participants who reported being suspicious. Unless indicated otherwise, all measured variables were assessed on seven-point scales, on which participants could indicate their level agreement (1 = *disagree completely*, 7 = *agree completely*). All participants provided informed consent and were debriefed, compensated, and thanked for their participation.

Experiment 1

In Experiment 1, we tested our hypotheses in an experimental tax-paying game. More specifically, we devised an experimental game in which an authority introduces a fine and provides no justification or justifies this fine as a means to deter. Participants could misreport their revenue to the authority to evade a stipulated tax rule.

Method

Participants and design. Seventy U.S. college students (62 females; M_{age} = 18.87 years, SD_{age} = 1.36) were randomly assigned to one of two justification conditions (deterrence vs. no justification). Participants received course credit for their participation.

Procedure.

Rule compliance. Consistent with previous work on experimental tax games (Bilotkach, 2006), participants were told that they were randomly assigned to be a "worker" in a work team consisting of eight members, while one other group member was randomly assigned to be the team leader. Workers could earn extra money by finding correct words among scrambled letters. The rule was that 40% of the money would have to be paid to the group leader. The team leader had to evenly redistribute this money among all team members such that all group members could share

in a part of total revenues (i.e., similar to taxes that have to be paid to a governmental authority).

Although participants were told that the money they would earn was contingent on the number of words they found (and could thus vary across participants, depending on their productivity), all participants received $1.50—regardless of the amount of words correctly noted. Crucially, workers then had to *self-report* the amount of money they had earned to the team leader. Participants were told that the team leader was able to verify if this self-reported amount was correct for only two workers (i.e., partial monitoring from an authority). As such, participants had the possibility to misreport the amount of money they earned, and keep more of their own revenue (analogous to social dilemma games; see Molenmaker, De Kwaadsteniet, & van Dijk, 2014). Thus, participants who reported $1.50 to the team leader fully complied with the rules, whereas lower self-reported amounts reflected less rule compliance.

Sanction justifications. Participants were informed that the team leader had the ability to fine those who were caught misreporting their revenue; the team leader could do this by decreasing the money participants earned with the task by $1. The type and severity of the sanction was thus held constant across conditions. In the deterrence condition, the team leader justified this sanction as a means to deter workers from misreporting their revenues. That is, the leader stated, "the primary aim of the fine is to deter workers from misreporting revenue." In the no justification condition, no justification was given (i.e., no additional information was provided by the team leader).[1]

Distrust. Perceived distrust was measured with the following three items, "I feel distrusted by the team leader," "I feel like the team leader does not trust me," and "I feel like the team leader assumes I am going to lie" (Cronbach's $\alpha = .78$).

Results

Distrust. Participants felt more distrusted in the deterrence condition ($M = 4.03$, $SD = 1.12$) compared with the no-justification condition ($M = 3.18$, $SD = 1.28$), $t(70) = 2.95$, $p = .004$, Cohen's $d = 0.70$.

Rule compliance. To test to what extent participants misreported revenue on average (thus regardless of condition), we compared the mean of money participants reported earning to the mean of $1.50 that represented full rule compliance. Please note that participants reported their earnings in dollar cents. On average, participants underreported the amount of money they had earned ($M = 91.81$, $SD = 31.37$), $t(70) = 15.52$, $p < .001$, $d = 3.27$. Participants' misreporting also depended on the sanction-justification manipulation. Participants were more likely to underreport the amount of money they earned in the deterrence condition ($M = 81.49$, $SD = 34.83$) compared with the no-justification condition ($M = 102.14$, $SD = 23.77$), $t(70) = 2.89$, $p = .005$, $d = 0.70$.

Mediation analyses. Feeling distrusted was negatively correlated with rule compliance, $r = -.38$, $p = .001$ and mediated the effect of the deterrence justification on rule compliance (95% CI = $[-15.26, -1.26]$, $\kappa^2 = .10$).

Discussion

Consistent with our hypotheses, participants were more likely to underreport their earnings to their team leader when this leader justified a sanction as a means to deter lying. This was explained by deterrence justifications increasing the degree to which participants felt distrusted by the team leader.

Experiment 2

In Experiment 2, we attempted to replicate the effect we observed in Experiment 1 while adding a just-deserts justification condition. This allowed us to demonstrate that the findings in Experiment 1 are specific to deterrence justifications (and not simply attributable to an authority providing additional sanction-goal information), and stay consistent with the previous literature on sanction goals that pits deterrence against just deserts (Carlsmith et al., 2002; Hobbes, 1651/1988; Kant, 1780/1961; Keller et al., 2010; Mooijman et al., 2015; Nagin, 1998).

Method

Participants and design. Three hundred twenty-six participants (198 males; $M_{age} = 34.73$ years, $SD_{age} = 10.91$) were recruited from the Mechanical Turk website and participated in exchange for $0.50. Participants were randomly assigned to one of three justification conditions (deterrence vs. just deserts vs. no justification).

Procedure.

Rule compliance. The experimental game used was identical as the game used in Experiment 1.

Sanction justifications. In the deterrence condition, the team leader justified this sanction as a means to deter workers from misreporting their revenue. The team leader stated, "the primary aim of the fine is to prevent workers from misreporting revenue." In the just deserts condition, the team leader justified the sanction as giving team members who misreport their revenue their just deserts. The team leader stated, "the primary aim of the fine is to give team members who misreport revenue their just deserts." In the no-justification condition, no justification was given (i.e., no additional information was provided by the leader).[1]

Distrust. Perceived distrust was measured with the following six-items, "I feel distrusted by the team leader," "I feel like the

[1] We conducted an additional study on Mechanical Turk ($N = 142$) to test what participants believed was the authority's motivation in the deterrence, just deserts, deterrence/just deserts, and control condition. We used the experimental tax game from Experiments 1 and 2 and included a deterrence, just deserts, deterrence/just deserts (see Experiment 4), and no-justification condition. We then asked participants whether they thought the authority was motivated to deter, provide just deserts, deter *and provide* just deserts, or whether the participants did not know the authority's motivation. Results provided strong evidence for the notion that the sanction-justification manipulations worked as intended; 98% of participants in the deterrence condition indicated that the authority was motivated to deter and 89% of participants in the just-deserts condition indicated that the authority was motivated to provide just deserts. Furthermore, 40% of participants in the control condition indicated that the authority was motivated to provide just deserts; 8% of participants indicated that the authority was motivated to deter *and* provide just deserts; 4% of participants indicated that the authority was motivated to deter; and 48% of participants in the control condition indicated that they did not know the authority's motivation. Lastly, 91% of participants in the deterrence/just-deserts condition indicated that the authority was motivated to deter *and* provide just deserts. These results demonstrate that the manipulations that are used are effective, and that deterrence tends not to be perceived as the authority's main motivation in the control condition.

team leader does not trust me," "I think the leader assumes I am going to lie," "I think the leader assumes I am going to break the rules," "The leader expects me to have bad intentions," and "I think the leader believes I am going to lie" ($\alpha = .83$).[2]

Results

Distrust. Overall, the sanction justification affected feelings of distrust, $F(1, 323) = 18.11$, $p < .001$, $\eta_p^2 = .10$. Participants felt more distrusted in the deterrence condition ($M = 4.83$, $SD = 1.36$) than in the just-deserts condition ($M = 4.11$, $SD = 1.29$), $t(217) = 4.04$, $p < .001$, $d = 0.55$, and no-justification condition ($M = 3.78$, $SD = 1.29$), $t(215) = 5.82$, $p < .001$, $d = 0.79$. Participants felt (marginal significantly) more distrusted in the just-deserts condition compared with the no-justification condition, $t(214) = 1.84$, $p = .067$, $d = 0.25$.

Rule compliance. Participants on average underreported their revenues to the team leader ($M = 92.41$, $SD = 57.63$), $t(325) = 18.04$, $p < .001$, $d = 2.00$. Participants' underreporting also depended on the sanction justification, $F(1, 323) = 4.29$, $p = .014$, $\eta_p^2 = .03$. Participants were more likely to underreport the amount of money they had earned in the deterrence condition ($M = 79.95$, $SD = 57.69$) compared with the just-deserts condition ($M = 101.93$, $SD = 57.38$), $t(217) = 2.83$, $p = .005$, $d = 0.39$, and no-justification condition ($M = 95.51$, $SD = 56.08$), $t(215) = 2.04$, $p = .045$, $d = 0.27$. Reporting behavior did not differ between the just-deserts condition and no-justification condition, $t(214) = 0.83$, $p = .41$, $d = 0.05$.

Mediation analyses. Feeling distrusted was negatively correlated with rule compliance ($r = -.15$, $p = .007$) and mediated the overall effect of justification condition on rule compliance (95% CI = [−4.42, −0.24], $\kappa^2 = .03$). This was the case for both the deterrence versus just-deserts contrast (95% CI = [−7.24, −1.05], $\kappa^2 = .03$) and the deterrence versus no-justification contrast (95% CI = [−5.83, −0.44], $\kappa^2 = .04$).

Discussion

Replicating Experiment 1, participants were more likely to underreport their revenues when the team leader justified a sanction as a means to deter lying compared with when the team leader provided no justification. Participants were also more likely to underreport their revenues when the team leader justified a sanction as a means to deter lying compared with as a means to give rule breakers their just deserts. Consistent with our hypotheses and with the results from Experiment 1, these effects were explained by participants feeling distrusted by the leader. These findings provide support for our hypotheses in a different sample, while also demonstrating that the observed findings are specific to deterrence (and not just deserts) justifications.

Experiment 3

In Experiment 3, we aimed to generalize our findings from experimental tax games to sanctions that are justified as a means to deter people from committing tax fraud. Moreover, Experiment 2 used the deterrence or just-deserts sanction justifications as mutually exclusive—that is, a sanction was justified as either aimed at deterrence or just deserts. In reality, however, these goals can be, and often are, combined. One could argue that just deserts might attenuate the negative relationship between deterrence and rule compliance (because just deserts also signals authorities' focus on past rule breakers). Experiment 3 aims to address this issue by investigating the effect of providing simultaneously a deterrence *and* just deserts justification. Because we theorize that deterrence justifications signal distrust by casting a wide net of suspicion that includes the participant, and that this distrust negatively affects rule compliance, we predicted that the presence of a deterrence justification negatively affects rule compliance regardless of whether it is provided simultaneously with a just deserts justification.

To rule out competing explanations, we also measured the extent to which the deterrence justification affects people's distrust toward others. Besides making people feel distrusted, authorities that provide deterrence justifications may also make people distrust other group members (because an authority also tries to deter them). People's distrust toward others has been shown to undermine cooperation through raising fears of exploitation by these others (Mulder et al., 2006). We predict that deterrence impacts rule compliance through increasing the extent to which people feel distrusted by an authority, even when controlling for people's distrust toward others. Deterrence, in other words, impacts rule compliance primarily through increasing the extent to which people feel distrusted by the authority.

Method

Participants and design. One hundred eight-six participants (113 males; $M_{age} = 33.70$ years, $SD_{age} = 10.42$) were recruited from the Mechanical Turk website and randomly assigned to one of three justification conditions (deterrence vs. deterrence/just deserts vs. no justification). Participants received $1.00 for their participation.

Procedure.

Sanction justifications. Participants read an excerpt on tax fraud. Specifically, they read that American citizens are sometimes tempted to commit tax fraud by, for instance, underreporting their earnings to the U.S. Internal Revenue Service (IRS). Participants were further informed that the IRS could fine citizens for committing tax fraud (no-justification condition). It was further added that the primary aim of IRS policies is to "deter citizens from committing such tax fraud with these sanctions," (deterrence condition) or both "deter citizens from committing tax fraud and give those who commit tax fraud their just deserts" (deterrence/just-deserts condition). In the deterrence/just deserts condition, the order of the deterrence and just deserts explanation was counterbalanced but this had no significant impact on the results.

Distrust. Feelings of distrust were measured on a three-item scale. Items included, "I feel distrusted by the IRS," "I think the IRS assumes I want to break tax rules," and "I feel like the IRS assumes I have bad intentions" ($\alpha = .96$).

[2] We also included a measurement of perceived authority anger (e.g., "I feel like the leader will be angry when subordinates lie; Wubben et al., 2011) to explain a possible effect of just deserts on rule compliance. However, we did not observe an effect of just deserts on rule compliance. In addition, controlling for perceived authority anger did not significantly change the results of deterrence on distrust and rule compliance.

Rule compliance. We adapted rule compliance measurements from previous research (see Tyler & Blader, 2005). Rule compliance was measured on a five-item scale. Items included, "I feel inclined to behave according to all rules set by the IRS," "I feel obliged to stick to the rules regarding tax fraud," "I will act according to the rules even when the IRS will never know if I committed tax fraud," and "I feel inclined to commit tax fraud when I can get away with it" (reverse-coded; $\alpha = .93$).

Distrust toward others. Consistent with Mulder et al. (2006), distrust toward other taxpayers was measured on a three-item scale. Items included, "I feel like I cannot trust other citizens to pay their taxes," "I think taxpayers are tempted to break the rules," and "I feel like taxpayers cannot be trusted," ($\alpha = .94$).

Results

The means and standard deviations are reported in Table 1.

Distrust. Overall, the sanction justification manipulation affected feelings of distrust, $F(1, 183) = 8.74$, $p < .001$, $\eta_p^2 = .09$. Participants felt more distrusted in the deterrence condition than in the no-justification condition, $t(123) = 4.09$, $p < .001$, $d = 0.74$. Distrust did not differ between the deterrence and deterrence/just-deserts condition, $t(121) = 1.02$, $p = .31$, $d = 0.18$. Moreover, distrust was higher in the deterrence/just-deserts condition than in the no-justification condition, $t(122) = 3.01$, $p = .003$, $d = 0.55$.

Rule compliance. The sanction justification manipulation also influenced the willingness to comply with IRS rules, $F(1, 245) = 4.93$, $p = .008$, $\eta_p^2 = .05$. Rule compliance did not differ between the deterrence condition and deterrence/just-deserts condition, $t(121) = 0.05$, $p = .96$, $d = 0.01$, but was lower in the deterrence condition and deterrence/just-deserts condition compared with the no-justification condition, $t(123) = 2.69$, $p = .008$, $d = 0.49$; $t(122) = 2.75$, $p = .007$, $d = 0.49$, respectively.

Distrust toward others. The sanction justification manipulation did not affect distrust toward others, $F(1, 183) = 0.38$, $p = .68$, $\eta_p^2 = .09$.

Mediation analyses. Using a bootstrap analysis procedure with 5,000 resamples (Hayes, Preacher, & Myers, 2011), we tested whether distrust toward others, or perceived distrust toward self mediated the effect of the sanction-justification manipulation on rule compliance. Results from the bootstrap analyses demonstrated that distrust toward others was correlated negatively with rule compliance ($r = -.21$, $p = .004$) but did not mediate the overall effect (or the contrast effects between the significant conditions) of sanction justifications on rule compliance (all 95% CIs fell between -0.20 and 0.09, without zero in the interval). Instead, results from the bootstrap analyses demonstrated that feeling distrusted by the IRS correlated negatively with rule compliance ($r = -.34$, $p < .001$) and mediated the overall effect of the sanction-justification manipulation on rule compliance (95% CI = [-0.17, $-.02$], $\kappa^2 = .07$), even after controlling for participants' distrust toward others (95% CI = [-0.15, $-.02$]). This was similar for the deterrence versus no-justification contrast (95% CI = [-0.79, -0.21], $\kappa^2 = .15$), and deterrence/just-deserts condition versus no-justification contrast (95% CI = [-0.22, $-.04$], $\kappa^2 = .11$).

Discussion

Experiment 3 replicates Experiments 1 and 2 in a different context, while demonstrating that a deterrence justification also undermines rule compliance when presented in combination with a just-deserts justification. Consistent with our theorizing, these effects were attributable to a deterrence justification increasing feelings of distrust and not increasing distrust toward others. These results strongly suggest that the rule undermining effect of sanction justifications is specific to deterrence.

Experiment 4

Experiment 4 extends the previous three experiments in two ways. First, we manipulated whether or not to a university justified its reliance on a sanctioning system as a means to deter plagiarism from students. This allowed us to generalize our findings from the fine used in the experimental tax games of Experiments 1 and 2 and the generic IRS sanctions in Experiment 3 to a frequently used and well-known sanction at universities (i.e., exclusion from a course). Second, we investigated the moderating role of the perceived legitimacy of an authority. Legitimacy is the belief that authorities have the right to govern and that people should comply with their rules (Tyler, 2006). Therefore, we predict that a deterrence justification increases distrust independent of legitimacy beliefs (i.e., because the authority stills communicates distrust to all potential rule breakers), but we predict that the extent to which this distrust undermines rule compliance is attenuated by perceptions of (high) legitimacy (i.e., because the belief that one should comply with rules should override the negative impact of feeling distrusted). Thus, we predict that distrust mediates the effect of deterrence on rule compliance to a greater extent when legitimacy is low compared with high. These findings are consistent with our theorizing on distrust as a perception of interpersonal (mis)treatment; legitimacy has been shown to act as a buffer against relational threats (Tyler, 2006).

Method

Participants and design. One hundred sixteen U.S. college students (79 females; $M_{age} = 23.42$ years, $SD_{age} = 6.64$) partici-

Table 1
Means and (Standard Deviations) as a Function of Justification Condition for Experiment 3

Variable	Justification conditions			
	Deterrence	Deterrence/Just deserts	No justification	Total
Feeling distrusted	4.57 (2.04)	4.20 (2.02)	3.14 (1.85)	3.97 (2.05)
Distrust others	3.28 (1.93)	3.38 (2.00)	3.09 (1.76)	3.26 (1.89)
Rule compliance	4.86 (1.63)	4.88 (1.60)	4.09 (1.59)	4.61 (1.64)

pated in exchange for course credit and were randomly assigned to one of two justification conditions (deterrence vs. no justification).

Procedure.

Legitimacy. Legitimacy was measured with the following six items, "The university has the right to sanction students," "The university always has the right to enforce rules," "The university always has the right to make students comply with the rules," "Decisions made by the university are legitimate," "Decisions made by the university should be accepted by students," and "The university has the right to exclude students from courses" (α = .91).

Sanction justification. All participants were informed that the study assessed students' attitudes toward university policy. More specifically, it was explained how university students might sometimes be tempted to directly copy information from professional articles for their own work. It was explained that the policy of their university was to immediately exclude students who committed such plagiarism from their respective courses. In the no-justification condition, no further information was given. In the deterrence condition, participants read, "the primary aim of this punitive policy is to deter students from committing plagiarism." The severity of the sanction was thus held constant across the two justification conditions.[1]

Distrust. Perceived distrust was measured with the following six items, "I feel distrusted by the university," "I feel like the university does not trust me," "I think the university assumes I am going to commit plagiarism," "I think the university assumes I am going to break the rules," "The university expects me to have bad intentions," and "I think the university believes I am going to break the rules" (α = .86).

Rule compliance. Rule compliance was assessed on an eight-item scale. This scale measured students' willingness to commit plagiarism. Items included, "I feel inclined to behave according to university rules regarding plagiarism," "I feel obliged to stick to the rules regarding plagiarism," "I will act according to the rules even when the university will never know if I committed plagiarism," and "I feel inclined to break plagiarism rules when I can get away with it" (reverse-coded; α = .75).

Results

Distrust. Multiple regression analysis was used to test the interactive effects of deterrence and legitimacy on distrust. For the first step, deterrence (coded as −1 for no justification and +1 for deterrence) and legitimacy (standardized) were included as predictors. For the second step, the interaction between deterrence and legitimacy was added. Results demonstrated main effects of deterrence and legitimacy (β = .55), $t(116)$ = 7.39, $p < .001$; (β = .23), $t(116)$ = 3.06, p = .003, but no interaction effect between deterrence and legitimacy (β = .21), $t(116)$ = 1.29, p = .20.

Rule compliance. Multiple regression analysis was used to test the interactive effects of deterrence and legitimacy on rule compliance. For the first step, deterrence and legitimacy (standardized) were included. For the second step, the interaction between deterrence and legitimacy was added. Results demonstrated main effects of deterrence and legitimacy (β = −.15), $t(116)$ = 1.19, p = .095; (β = .36), $t(116)$ = 4.22, $p < .001$, and an interaction effect between deterrence and legitimacy (β = −.39), $t(116)$ = −2.08, p = .040. Results from a similar analysis including distrust, legitimacy, and their interaction demonstrated a main effect of legitimacy (β = .33), $t(116)$ = 4.18, $p < .001$, no main effect of distrust (β = −.08), $t(116)$ = 1.06, p = .29, and a significant interaction effect between distrust and legitimacy (β = .20), $t(116)$ = 3.02, p = .003.

A similar multiple regression analysis including deterrence, distrust, legitimacy and both interaction terms (deterrence/legitimacy, distrust/legitimacy), yielded a significant interaction term between distrust and legitimacy (β = −.19), $t(116)$ = −2.11, p = .037, but no significant interaction term for deterrence and legitimacy (β = −.11), $t(116)$ = −1.16, p = .11.

This demonstrates that legitimacy moderates the relationship between distrust and rule compliance instead of the relationship between deterrence and rule compliance (Aiken & West, 1991). Follow-up analyses demonstrated that distrust only significantly (and negatively) predicted rule compliance when legitimacy was relatively low (β = −.26), $t(116)$ = −2.71, p = .007, but not when legitimacy was relatively high (β = .10), $t(116)$ = 1.03, p = .30 (see Figure 1).

Mediation analysis. Replicating the previous three experiments, feeling distrusted was negatively correlated with rule compliance (r = −.25, $p < .001$) and mediated the main effect of the deterrence justification on rule compliance (95% CI = [−0.22, −0.05], κ^2 = .11). Moreover, we used Model 14 with 5,000 bootstraps in PROCESS (Hayes, 2013) to test the degree to which distrust mediated the relationship between deterrence and rule compliance when we allow legitimacy to moderate the relationship between distrust and rule compliance (as demonstrated earlier). Results yielded a significant moderated mediation effect (95% CI = [−0.14, −0.01]); distrust mediated the negative relationship between deterrence and rule compliance only when legitimacy was low (−1 *SD*; indirect effect = 0.12, *SE* = 0.06, 95% CI = [0.04, 0.24]), but not when legitimacy was high (+1 *SD*; indirect effect = −0.02, *SE* = 0.06, 95% CI = [−0.12, 0.09]; see Figure 2).

Discussion

These results extend the findings from the previous three experiments in at least two ways. First, we replicate the main effect of the deterrence justification in a different context that used sanctions that are more commonly encountered (i.e., a university sanction system). Second, we demonstrate that legitimacy attenuated the degree to which the distrust that was fostered by the deterrence justification impacted rule compliance. Indeed, the de-

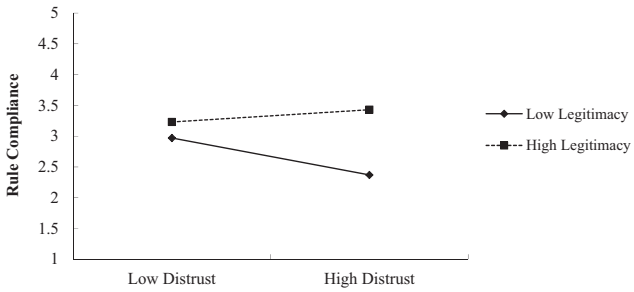

Figure 1. Interactive effects of legitimacy and distrust on rule compliance for Experiment 4.

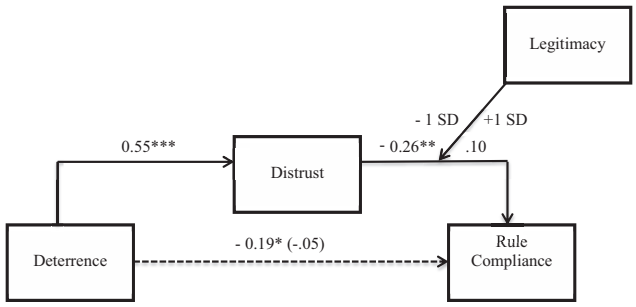

Figure 2. Mediation analysis for Experiment 4. Deterrence is coded as 1, control as −1. Beta weights are standardized. * $p < .05$. ** $p < .01$. *** $p < .001$.

terrence justification increased distrust regardless of perceived authority legitimacy (consistent with the notion that the justification communicates distrust to all potential rule breakers, including participants), but this distrust only negatively affected rule compliance when perceived legitimacy was relatively low, not when it was relatively high. These findings are consistent with our theorizing on distrust as a perception of interpersonal (mis)treatment; legitimacy has been shown to act as a buffer against relational threats (see Tyler, 2006). Legitimacy, in other words, can "override" the impact of feeling distrusted on rule compliance.

Experiment 5

Experiment 4 identified an important variable that can counteract the impact of distrust on rule compliance. However, legitimacy is not always easy to gain (people's perception of authorities are unlikely to change overnight) and Experiment 4 did not provide unequivocal support for legitimacy attenuating the relationship between deterrence and rule compliance. Understanding how authorities can directly prevent the negative effects of a deterrence justification in a more practical manner seems desirable. We therefore investigated the effects of framing a deterrence justification as aimed at a group of people that either did or did not include the participant. If a deterrence justification fosters distrust through signaling that one is considered a potential rule breaker, then a deterrence justification aimed at *oneself* should increase distrust, whereas a deterrence justification aimed *at others* should attenuate the extent to which one feels distrusted. We predicted that a sanction signals more distrust (and is thus less effective in promoting a willingness to comply with rules) when it is justified as deterring oneself—compared with others—from rule breaking, and compared with a sanction that is provided without a justification. In addition, we investigated people's attitudes toward the authority. Liking of an authority can be an important predictor of people's willingness to comply with authorities' rules (Tyler & Blader, 2000) and authorities that provide deterrence justifications may be liked less because they make people feel distrusted. We therefore (a) controlled for attitudes toward the authority to demonstrate that the impact of deterrence on rule compliance cannot be fully attributed to authority liking, and (b) investigated whether feeling distrusted by an authority undermined rule compliance through increasing negative attitudes toward the authority.

Method

Participants and design. One hundred eighty-five U.S. participants (110 males; $M_{age} = 33.19$ years, $SD_{age} = 11.09$) were recruited from the Mechanical Turk website and were randomly assigned to one of three justification conditions (self deterrence vs. other deterrence vs. no justification). Participants received $1 for their participation.

Procedure.

Rule compliance. The experimental game used was similar to the game used in Experiments 1 and 2, except for two differences. First, instead of self-reporting to the team leader how much money they received, participants were first informed that according to the rules set by the team leader, their work corresponded to a $1 reward. Participants could then each take up to $7 from the team leader as a reward for their own performance. The team leader was able to verify if the money taken was the correct amount ($1) for only two workers (i.e., partial monitoring from an authority). As such, participants had the possibility to take more money than they had earned. Second, the amount of money participants could gain was thus higher than in Experiments 1 and 2. That is, seven times as much as they received for simply participating in the study.

Sanction justifications. It was explained to participants (who all reported to be U.S. citizens) that Mechanical Turk workers hail from different countries; participants read that Mechanical Turk workers from different countries have been shown to behave differently in studies that revolve around money. The justification conditions were identical to Experiment 1 with the exception that the fine was justified by the leader as a means to deter U.S. participants (deterrence self condition) versus non-U.S. participants (deterrence other condition) from taking too much money from the team leader.[1]

Distrust. Perceived distrust was measured with the following four items, "I feel distrusted by the team leader," "I feel like the team leader does not trust me," "I think the leader assumes I am going to lie," "I think the leader assumes I am going to break the rules," ($\alpha = .83$).

Attitudes toward supervisor. We measured participants' attitudes toward the supervisor on four items. Items included, "I like this team leader," "I have a positive feeling about the team leader," "I tend to view this team leader positively," and "I dislike the team leader" (reverse-coded; $\alpha = .93$)

Results

Distrust. The sanction justification manipulation affected the extent to which participants felt distrusted by the team leader, $F(1, 182) = 19.31$, $p < .001$, $\eta_p^2 = .18$. Perceived distrust was higher in the deterrence self condition ($M = 4.57$, $SD = 1.83$) compared with the deterrence other condition ($M = 2.90$, $SD = 1.69$), $t(120) = 5.36$, $p < .001$, $d = 0.99$, and no-justification condition ($M = 2.79$, $SD = 1.87$), $t(120) = 5.32$, $p < .001$, $d = 0.97$. The deterrence other condition did not differ from the no-justification condition, $t(120) = 0.29$, $p = .77$, $d = 0.04$.

Rule compliance. On average, participants took more money than they earned ($M = 26.62$, $SD = 18.96$), $t(184) = 9.93$, $p < .001$, $d = 1.46$. Participants' rule compliance also depended on the sanction-justification manipulation. The sanction-justification manipulation influenced rule compliance, $F(1, 182) = 4.38$, $p = .014$, $\eta_p^2 = .05$. Participants took more money in the deterrence self

condition ($M = 31.70$, $SD = 20.99$) compared with the deterrence other condition ($M = 24.00$, $SD = 17.92$), $t(124) = 2.24$, $p = .037$, $d = 0.38$, and no-justification condition ($M = 22.50$, $SD = 16.56$), $t(120) = 2.67$, $p = .008$, $d = 0.47$. The deterrence other condition and no-justification condition did not differ, $t(120) = 0.46$, $p = .65$, $d = 0.07$.

Attitudes toward authority. The sanction-justification manipulation influenced participants' attitudes toward the supervisor, $F(1, 241) = 10.71$, $p < .001$, $\eta_p^2 = .12$. Participants in the deterrence self condition liked the supervisor less ($M = 3.79$, $SD = 1.46$) than participants in the deterrence other condition ($M = 4.46$, $SD = 1.53$), $t(124) = 2.52$, $p = .013$, $d = 0.46$, and no-justification condition ($M = 4.86$, $SD = 1.61$), $t(120) = 3.85$, $p < .001$, $d = 0.70$. The deterrence other condition and no-justification condition did not differ, $t(120) = 1.40$, $p = .16$, $d = 0.25$.

Mediation analyses. Using a bootstrap analysis procedure with 5,000 resamples (Hayes et al., 2011), we tested whether distrust mediated the effect of the sanction-justification manipulation on rule compliance. Distrust was positively correlated with taking more money than participants earned ($r = .62$, $p < .001$), and results from the bootstrap analysis showed that distrust mediated the overall effect of the sanction-justification manipulation on rule compliance (95% CI = [−7.92, −3.40], $\kappa^2 = .26$), even after controlling for attitudes toward the supervisor (95% CI = [−5.41, −1.69]) or after adding attitudes toward the supervisor as an additional mediator (95% CI = [−6.93, −2.59]). Attitudes toward the supervisor also independently mediated the effect of the sanction-justification manipulation on rule compliance (95% CI = [−2.15, −0.14]). The indirect effect of distrust was also significant for the deterrence self versus no-justification contrast (95% CI = [−16.36, −5.82) and the deterrence self versus deterrence other contrast (95% CI = [−9.01, −3.79]). Interestingly, for both contrasts the deterrence justification undermined rule compliance through the mediating effect of distrust predicting negative attitudes toward the authority (i.e., deterrence→distrust→attitudes toward authority→rule compliance; 95% CIs: [−6.93, −2.59]).

Discussion

The findings from Experiment 5 demonstrate that the extent to which a deterrence justification fostered distrust and decreased sanction effectiveness was attenuated when it was explicitly aimed at others. This corroborates our assumption that a deterrence justification fosters distrust through signaling that one is considered a potential rule breaker. Moreover, no differences in distrust were observed when the sanction was justified as deterring others compared with the no-justification condition. This strongly suggests that the distrust that authorities communicate with a deterrence justification is in part attributable to people inferring that the sanction is aimed at them. Lastly, the findings from Experiment 5 demonstrate that feeling distrusted by an authority undermines rule compliance in part through increasing negative attitudes toward this authority.

General Discussion

We presented five experiments that examined how sanction justifications affect sanction effectiveness. Across different samples, contexts, and sanctions, we consistently observed that people feel more distrusted when sanctions are justified as attempts to deter them from breaking rules. This distrust is shown to decrease people's willingness to comply with the rules of their team leader (Experiments 1, 2, and 5), IRS (Experiment 3), and university (Experiment 4). These results strongly suggest that justifying a sanction as an attempt to deter people from breaking rules makes people feel more distrusted (Hypothesis 1) which decreases the effectiveness of a sanction (Hypotheses 2a and 2b).

Theoretical and Practical Implications

The present set of studies makes several contributions to the literature on rule compliance. Previous research has mainly focused on the extent to which people use deterrence and just-deserts goals in guiding their sanction decisions (Carlsmith, 2006, 2008; Carlsmith et al., 2002; Darley et al., 2000; Gerber & Jackson, 2012; Mooijman et al., 2015). The present research is to our knowledge the first to demonstrate the effects of using one such goal (deterrence) as a *justification*. The reported studies demonstrate that using a deterrence goal as a justification can lead an authority to signal the sanction's underlying considerations to the public. Sanction goals are therefore not only "hidden" motivations that drive punitive-sanction decisions (Carlsmith et al., 2002); as justifications, they can also be relevant for influencing the effectiveness of sanctions. Interestingly, the present research demonstrates that deterrence goals do not only influence rule compliance through affecting the severity and type of sanction. Rather, deterrence goals can have a distinct and independent influence, regardless of sanction type and severity. The underlying considerations that an authority signals to others through deterrence justifications are therefore highly relevant for the subsequent effectiveness of the sanction.

The way authorities affect others' tendency to comply with the rules is complex. Sanctions are not just means to increase the costs and decrease the benefits of rule breaking (Nagin, 1998). Rather, sanctions are also driven by philosophies and goals (Bentham, 1789/1988; Hobbes, 1651/1988; Kant, 1780/1961) that can directly affect the public. Previous research has demonstrated that powerful authorities are inclined to rely on deterrence as a sanction goal because they distrust others (Mooijman et al., 2015). The current research implies that the public is able to infer this distrust from sanctions justified as attempts to deter rule breaking. This stresses the notion that authorities should be cautious with how they justify their sanctions. Indeed, the perceived distrust that was elicited by deterrence justifications played a unique role in undermining rule compliance. The experiments demonstrated that feeling distrusted seems to have at least a moderate degree of influence on rule compliance (Cohen, 1988; Preacher & Kelley, 2011) and this rule-undermining effect of feeling distrusted was independent from other variables such as perceived anger (Wubben et al., 2011), attitudes toward authorities (Tyler & Blader, 2003), and distrust toward others (Mulder et al., 2006). The current research thus demonstrates that perceived authority (dis)trust is of importance for an authority's ability to promote compliance with cooperative rules.

As such, the current studies have direct practical relevance for authorities—judges, policymakers, and managers should be aware of the consequences that a deterrence justification can have for

sanction effectiveness. To stimulate rule compliance it may perhaps be more effective to (a) emphasize that sanctions are meant to give people their just deserts, (b) give no sanction justification, or (c) use deterrence justifications that signal that the sanction targets others. Such sanctions foster rule compliance without making people feel (too) distrusted. Although this advice seems straightforward, it may be harder to achieve than one may think. Recent research has demonstrated that power increases people's reliance on deterrence as a goal for punishment (Mooijman et al., 2015); power increases distrust toward others, which increases reliance on deterrence as a goal for punishment. Powerful authorities may thus ironically be the most inclined to emphasize deterrence aspects of a sanction, even though this may be one of the least effective course of action.

The present research suggests, however, that deterrence justifications are more effective when they are coupled with an affirmation of trust in the target (e.g., sanctions are not targeted toward the self). This provides at least one way in which authorities can attenuate the effects of deterrence justifications. Moreover, the distrust that is fostered by deterrence justifications is shown to be less likely to impact rule compliance when authority legitimacy is high. Ironically, however, authorities with low legitimacy tend to suffer most from rule compliance problems (Tyler, 1990). This suggests that when authority legitimacy is low, deterrence justifications might exacerbate the rule-compliance problems of authorities (since their legitimacy is not attenuating the effects of people feeling distrusted).

Limitations and Future Directions

Whereas the present research supports our hypotheses across different samples, contexts, and measures, there are some issues to be noted. For instance, we did not focus on individuals who have already broken rules. First, it is possible that these individuals feel distrusted even though this distrust would be partially justified. Although potentially interesting, the majority of people tend to be rule abiding or at least perceive themselves as such (Brown, 2012; Sedikides, Meek, Alicke, & Taylor, 2014). Justifying a sanction as a deterrent may therefore still make rule breakers feel distrusted and thereby undermine sanction effectiveness. Second, it is also possible that although rule breakers feel distrusted by an authority that aims to deter them from breaking rules, this distrust may not translate into rule-breaking behavior (because by this they would provide the authority a legitimization for their distrust). Third, it is also possible that the moral self-image of participants moderates the degree to which distrust undermines rule compliance. Feeling distrusted may at times stimulate more rule compliance for individuals who are more likely to perceive themselves of moral individuals. This may especially be the case when individuals can explicitly demonstrate their morality to others (and thereby restore their moral self-image in their eyes and the eyes of others). Although we did not measure this in the present studies, we believe this to be an interesting avenue for future research. More research may also be needed to flesh out exactly why feeling distrusted undermines rule compliance. Although our findings are consistent with our theorizing and also demonstrate the role of attitudes toward the authority, future research could benefit from directly testing the notion that feeling distrusted is perceived as a form of interpersonal injustice. This could potentially explain why feeling distrusted by an authority, but not distrusting others (which is less likely to be perceived as an interpersonal injustice), explained the negative impact of deterrence justifications on rule compliance. Indeed, feeling distrusted by an authority may also decrease the extent to which people place trust in an authority, thereby contributing to a self-sustaining cycle of mutual distrust between people and authorities.

Moreover, sanction goals are typically classified as deterrence or just-deserts goals (Carlsmith et al., 2002). Previous research has primarily focused on the extent to which people use these two goals in guiding there sanction decisions (Carlsmith, 2006, 2008; Carlsmith et al., 2002; Darley et al., 2000; Gerber & Jackson, 2012). The present research has been consistent with this approach because we examined how these two goals affect sanction effectiveness. However, we did not find any effects of a just-deserts justification on rule compliance, mainly because just deserts did not significantly affect distrust. Future research could examine this issue further; for example, it could be that framing the wording of a just-deserts justification in terms of retribution or revenge is more effective in affecting rule compliance through—for instance—influencing the anger that people infer from the authority. The perceived link between revenge and anger may be stronger than the perceived link between just deserts and anger (just deserts revolves about proportionality between crime and punishment, whereas revenge revolves around doling out disproportionate punishments to humiliate the other; see Gerber & Jackson, 2013); the link between revenge and anger is indeed well documented (Eisenberger, Lynch, Aselage, & Rohdieck, 2004; Seip, van Dijk, & Rotteveel, 2014).

Future research could also examine how sanction severity interacts with sanction justifications; the present research mainly used relatively mild sanctions, or unspecified sanctions. For instance, the negative impact of deterrence justifications could be amplified when sanctions are perceived as severe (because it signals more distrust). Although the focus in the present article is on how authorities justify their sanctioning behavior, future research could also investigate how people infer sanction goals from sanctions. Although this lies outside of the scope of the present article, our theorizing and reported experiments can provide a meaningful theoretical framework for formulating predictions about sanction inferred goals. For instance, inferring that a sanction is meant to deter rule breaking can make people feel distrusted, and thus undermine rule compliance. Future research could aim to investigate when sanctions are perceived to reflect a deterrence (or just-deserts) goal. Lastly, the present research primarily measured rule compliance in relatively small-stakes behavioral experiments. This approach provides experimental control and provides the advantage of being able to establish causality—consistent with a large literature in experimental and behavioral economics—but lacks empirical validation in high-stakes situations. To address this issue, future research could investigate the effects of sanction-goal justifications in real-life field settings involving high-stakes situations (e.g., ethical behavior in government and/or corporate settings).

Conclusion

Authorities frequently use sanctions to promote rule compliance and often provide a justification for their use of such sanctions.

Although providing a deterrence justification may seem appealing, the present article demonstrates that this can in fact negatively influence how effective a sanction will be in promoting future rule compliance. Five experiments demonstrate that sanction effectiveness decreases when sanctions are justified as attempts to deter rule breaking. This can be explained by deterrence justifications fostering feelings of distrust. This suggests that although authorities have been shown to rely on deterrence as a sanction justification (Mooijman et al., 2015), this may—paradoxically—undermine the effectiveness of sanctions.

References

Aiken, L., & West, S. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: SAGE.

Ball, G. A., Treviño, L. K., & Sims, H. P., Jr. (1994). Just and unjust punishment: Influences on subordinate performance. *Academy of Management Journal, 37,* 299–322. http://dx.doi.org/10.2307/256831

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137,* 594–615. http://dx.doi.org/10.1037/a0023489

Bentham, J. (1988). *An introduction to the principles of morals and legislation*. New York, NY: Prometheus Books. (Original work published 1789)

Bilotkach, V. (2006). A tax evasion-bribery game: Experimental evidence from Ukraine. *The European Journal of Comparative Economics, 3,* 31–49.

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin, 38,* 209–219. http://dx.doi.org/10.1177/0146167211432763

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42,* 437–451. http://dx.doi.org/10.1016/j.jesp.2005.06.007

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research, 21,* 119–137. http://dx.doi.org/10.1007/s11211-008-0068-x

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83,* 284–299. http://dx.doi.org/10.1037/0022-3514.83.2.284

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Cropanzano, R., Bowen, D. E., & Gilliand, S. W. (2007). The management of organizational justice. *The Academy of Management Perspectives, 21,* 34–48. http://dx.doi.org/10.5465/AMP.2007.27895338

Darley, J. M. (2009). Morality in the law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science, 5,* 1–23. http://dx.doi.org/10.1146/annurev.lawsocsci.4.110707.172335

Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. *Law and Human Behavior, 24,* 659–683. http://dx.doi.org/10.1023/A:1005552203727

De Cremer, D., & van Knippenberg, D. (2002). How do leaders promote cooperation? The effects of charisma and procedural fairness. *Journal of Applied Psychology, 87,* 858–866. http://dx.doi.org/10.1037/0021-9010.87.5.858

Eisenberger, R., Lynch, P., Aselage, J., & Rohdieck, S. (2004). Who takes the most revenge? Individual differences in negative reciprocity norm endorsement. *Personality and Social Psychology Bulletin, 30,* 787–799. http://dx.doi.org/10.1177/0146167204264047

Ellemers, N. (2012). The group self. *Science, 336,* 848–852. http://dx.doi.org/10.1126/science.1220987

Ellemers, N., Doosje, B., & Spears, R. (2004). Sources of respect: Effects of being liked by ingroups and outgroups. *European Journal of Social Psychology, 34,* 155–172. http://dx.doi.org/10.1002/ejsp.196

Fiske, S. T. (1993). Controlling other people. The impact of power on stereotyping. *American Psychologist, 48,* 621–628. http://dx.doi.org/10.1037/0003-066X.48.6.621

Gerber, M. M., & Jackson, J. (2013). Retribution as revenge and retribution as just deserts. *Social Justice Research, 26,* 61–80. http://dx.doi.org/10.1007/s11211-012-0174-7

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York, NY: Guilford Press.

Hayes, A. F., Preacher, K. J., & Myers, T. A. (2011). Mediation and the estimation of indirect effects in political communication research. In E. P. Bucy & R. Lance Holbert (Eds.), *Sourcebook for political communication research: Methods, measures, and analytical techniques* (pp. 434–465). New York, NY: Routledge.

Hobbes, T. (1996). *Leviathan*. New York, NY: Oxford University Press. (Original work published 1651)

Inesi, E. M., Gruenfeld, D. H., & Galinsky, A. D. (2012). How power corrupts relationships: Cynical attributions for others' generous acts. *Journal of Experimental Social Psychology, 48,* 795–803. http://dx.doi.org/10.1016/j.jesp.2012.01.008

Kant, I. (1952). The science of right. In R. Hutchins (Ed.), *Great books of the western world* (W. Hastie, Trans.) (pp. 397–446). Edinburgh, Scotland: T. & T. Clark. (Original work published 1790)

Keller, L. B., Oswald, M. E., Stucki, I., & Gollwitzer, M. (2010). A closer look at an eye for an eye: Layperson's punishment decisions are primarily driven by retributive motives. *Social Justice Research, 23,* 99–116. http://dx.doi.org/10.1007/s11211-010-0113-4

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review, 110,* 265–284. http://dx.doi.org/10.1037/0033-295X.110.2.265

Kirchler, E., Kogler, C., & Muelbacher, S. (2014). Cooperative tax compliance: From deterrence to deference. *Current Directions in Psychological Science, 23,* 87–92. http://dx.doi.org/10.1177/0963721413516975

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50,* 569–598. http://dx.doi.org/10.1146/annurev.psych.50.1.569

Langlois, J. (2012). *India to name and shame rapists. Global Post*. Retrieved from http://www.globalpost.com/dispatch/news/regions/asiapacific/india/121227/india-name-and-shame-rapists

Martinez, M. (2015, January 24). Colorado woman gets 4 years for wanting to join ISIS. *CNN*. Retrieved, January 28 from http://www.cnn.com/2015/01/23/us/colorado-woman-isis-sentencing/

McKenzie, C. R. M., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science, 17,* 414–420. http://dx.doi.org/10.1111/j.1467-9280.2006.01721.x

Molenmaker, W. E., De Kwaadsteniet, E. W., & Van Dijk, E. (2014). On the willingness to costly reward cooperation and punish non-cooperation: The moderating role of type of social dilemma. *Organizational Behavior and Human Decision Processes, 125,* 175–183. http://dx.doi.org/10.1016/j.obhdp.2014.09.005

Mooijman, M., van Dijk, W. W., Ellemers, N., & van Dijk, E. (2015). Why leaders punish: A power perspective. *Journal of Personality and Social Psychology, 109,* 75–89. http://dx.doi.org/10.1037/pspi0000021

Mulder, L. B., & Nelissen, R. (2010). When rules really make a difference: The effect of cooperation rules and self-sacrificing leadership on moral norms in social dilemmas. *Journal of Business Ethics, 95,* 57–72. http://dx.doi.org/10.1007/s10551-011-0795-z

Mulder, L. B., van Dijk, E., De Cremer, D., & Wilke, H. A. M. (2006). Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology, 42,* 147–162. http://dx.doi.org/10.1016/j.jesp.2005.03.002

Nagin, D. (1998). Criminal deterrence research at the outset of the twenty-first century. In M. Tonry (Ed.), *Crime and justice: A review of research* (Vol. 23, pp. 1–42). Chicago, IL: University of Chicago Press. http://dx.doi.org/10.1086/449268

Pelletier, L. G., & Vallerand, R. J. (1996). Supervisors' beliefs and subordinates' intrinsic motivation: A behavioral confirmation analysis. *Journal of Personality and Social Psychology, 71,* 331–340. http://dx.doi.org/10.1037/0022-3514.71.2.331

Pillutla, M., Malhotra, D., & Murnighan, J. K. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology, 39,* 448–455. http://dx.doi.org/10.1016/S0022-1031(03)00015-5

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16,* 93–115. http://dx.doi.org/10.1037/a0022658

Schilke, O., Reimann, M., & Cook, K. S. (2015). Power decreases trust in social exchange. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 112,* 12950–12955. http://dx.doi.org/10.1073/pnas.1517057112

Sedikides, C., Meek, R., Alicke, M. D., & Taylor, S. (2014). Behind bars but above the bar: Prisoners consider themselves more prosocial than non-prisoners. *British Journal of Social Psychology, 53,* 396–403. http://dx.doi.org/10.1111/bjso.12060

Seip, E. C., van Dijk, W. W., & Rotteveel, M. (2014). Anger motivates costly punishment of unfair behavior. *Motivation and Emotion, 38,* 578–588.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after p-hacking*. Paper presented at the Fourteenth Annual Meeting of the Society for Personality and Social Psychology, New Orleans, LA.

Snyder, M. (1992). Motivational foundations of behavioral confirmation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 67–113). New York, NY: Academic Press.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). New York, NY: Academic Press. http://dx.doi.org/10.1016/S0065-2601(08)60229-4

Tannenbaum, D., Valasek, C. J., Knowles, E. D., & Ditto, P. H. (2013). Incentivizing wellness in the workplace: Sticks (not carrots) send stigmatizing signals. *Psychological Science, 24,* 1512–1522. http://dx.doi.org/10.1177/0956797612474471

Tetlock, P. E., Visser, P., Singh, R., Polifroni, M., Elson, B., Mazzocco, P., & Rescober, P. (2007). People as intuitive prosecutors: The impact of social-control goals on attributions of responsibility. *Journal of Experimental Social Psychology, 43,* 195–209. http://dx.doi.org/10.1016/j.jesp.2006.02.009

Tyler, T. R. (1990). *Why we obey the law*. New Haven, CT: Yale University Press.

Tyler, T. R. (2006). Psychological perspectives on legitimacy and legitimation. *Annual Review of Psychology, 57,* 375–400. http://dx.doi.org/10.1146/annurev.psych.57.102904.190038

Tyler, T. R., & Blader, S. L. (2000). *Cooperation in groups: Procedural justice, social identity, and behavioral engagement*. Philadelphia, PA: Psychology Press.

Tyler, T. R., & Blader, S. L. (2003). The group engagement model: Procedural justice, social identity, and cooperative behavior. *Personality and Social Psychology Review, 7,* 349–361. http://dx.doi.org/10.1207/S15327957PSPR0704_07

Tyler, T. R., & Blader, S. L. (2005). Can businesses effectively regulate employee conduct? The antecedents of rule following in work settings. *Academy of Management Journal, 48,* 1143–1158. http://dx.doi.org/10.5465/AMJ.2005.19573114

Tyler, T. R., & Lind, E. A. (1992). A relational model of authority in groups. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 25, pp. 115–191). New York, NY: Oxford University Press. http://dx.doi.org/10.1016/S0065-2601(08)60283-X

Wubben, M. J. J., De Cremer, D., & van Dijk, E. (2011). The communication of anger and disappointment helps to establish cooperation through indirect reciprocity. *Journal of Economic Psychology, 32,* 489–501. http://dx.doi.org/10.1016/j.joep.2011.03.016

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion, 18,* 129–166. http://dx.doi.org/10.1007/BF02249397

Zand, D. (1997). *The leadership triad: Knowledge, trust, and power*. New York, NY: Oxford University Press.

Received May 21, 2015
Revision received October 18, 2016
Accepted October 19, 2016 ■