Judgment and Decision Making, Vol. 4, No. 7, December 2009, pp. 543–553

Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions

Rob M. A. Nelissen* and Marcel Zeelenberg
Tilburg University, the Netherlands

Abstract

Third-party punishment has recently received attention as an explanation for human altruism. Feelings of anger in response to norm violations are assumed to motivate third-party sanctions, yet there is only sparse and indirect support for this idea. We investigated the impact of both anger and guilt feelings on third-party sanctions. In two studies both emotions were independently manipulated. Results show that anger and guilt independently constitute sufficient but not necessary causes of punishment. Low levels of punishment are observed only when neither emotion is elicited. We discuss the implications of these findings for the functions of altruistic sanctions.

Keywords: third-party punishment, social norms, emotions, decision-making.

1 Introduction

People often defend the interests of others. They stand up for their friends if someone speaks ill about them in their absence. They do not tolerate a colleague being bullied at work. They boycott consumer products that are produced using child labor. Some even come to the aid of a stranger who is being physically harassed, in spite of obvious personal danger. In general, people retaliate against injustice even if they are not directly victimized. Sanctioning of norm-violations is vital for prosocial behavior to be sustained (Fehr & Gächter, 2002; Yamagishi, 1986). However, punishing norm-violations is costly in terms of time and energy. It may even impose physical risks. Punishing injustice is therefore considered to be a moral act, particularly when it is performed on behalf of others (i.e., in case of third-party sanctions, Fehr & Fishbacher, 2004). This begs the question of what incites third-party sanctions, as they usually oppose self-interest.

1.1 Moral emotions and prosocial behavior

Classic philosophical treatises on "moral sentiments" already stressed the functional role of moral emotions as elicitors of prosocial behavior (e.g., Hume, 1739; Smith, 1759). Moral emotions are defined as feelings related to the interest and welfare of others rather than one's own (Haidt, 2003). We experience feelings like empathy, anger, and guilt if we consider how others have been hurt, wronged, or harmed (e.g., Batson, 2006; Haidt, 2003). The view that moral emotions also have functional *behavioral* consequences is reflected in evolutionary hypothesis about their adaptive value (e.g., Tooby & Cosmides, 1990; Trivers, 1971).

Especially Robert Frank's (2004) notion of emotions as "commitment devices" seems relevant to understand how moral emotions promote prosocial behavior, in spite of the associated costs. Frank argues that moral emotions have been evolved as commitment devices that make people forego their immediate self-interest, committing them to a more rewarding long-term strategy. For instance, an angry individual retaliating a norm-violation may incur an immediate cost, but may derive a greater benefit in the long run by deterring future exploitation. Similarly, Frank also argued that guilt feelings act as a commitment device because a guilty person may invest time and energy to make up for something (s)he did to another person, but may eventually benefit thereof by saving a mutually rewarding and beneficial relationship. Precisely this effect was empirically supported (De Hooge, Breugelmans, and Zeelenberg, 2007; Ketelaar & Au, 2003; Nelissen, Dijker & de Vries, 2007). Moral emotions, particularly anger, have also been proposed as the proximal mechanism underlying third-party sanctions (Fehr & Fishbacher, 2004; Fehr & Gachter, 2002). Nevertheless, the proposed role for emotions in third-party punishment requires further exploration for two main reasons.

First of all, anger has not yet been linked empirically to *third-party* sanctions. The associations may seem straightforward, as numerous studies have related

*We thank Urs Fischbacher for providing us with the instructions for the third-party punishment paradigm, Johan Karremans for useful comment, and Jon Baron for his contributions. Correspondence can be addressed to Rob Nelissen, Department of Social Psychology and TIBER (Tilburg Institute of Behavioral Economics Research), Tilburg University, PO BOX 90153, 5000-LE Tilburg, the Netherlands, Email: r.m.a.nelissen@uvt.nl.

feelings of anger — either self-reported or at the physiological level — to retaliation of personal ill treatment (e.g., Ben-Shakhar, Bornstein, Hopfensitz, & Van Winden, 2007; Bosman & Van Winden, 2002; Pillutla & Murnighan, 1996; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Still, third-party sanctions are different, as perpetrators of norm-violations are punished not for what they did to the punisher, but for what they did to someone else. Recent insights clearly show that anger about personal harm is a distinct emotion from empathic anger at witnessing injustice or harm to someone else (Batson et al., 2007). Since both types of anger have different elicitors, it remains to be demonstrated whether (empathic) anger also instigates third-party punishment.

Second, moral emotions fall into two large categories (Haidt, 2003). On the one hand, other-focused emotions like anger arise in response to perceiving *someone else intentionally* causing harm to another person. On the other hand, self-focused emotions like guilt feelings arise in response to or anticipation of *oneself being responsible* for another person's misfortune. Several studies have shown that feelings of guilt are associated with prosocial behavior (e.g., De Hooge et al., 2007; Ketelaar & Au, 2003; Nelissen et al., 2007). Insofar as third-party punishment is a form of "second-order cooperation" (Fehr & Gächter, 2002; Yamagishi, 1986, 1988), it may be fuelled not only by anger over norm-violations, but also by anticipated guilt with respect to not punishing this violation when it would be one's responsibility to do so.

1.2 Anger, guilt, and third-party sanctions

We thus predict that third-party sanctions could be elicited not only by feelings of anger, but also by feelings of guilt. However, even though both emotions may underlie third-party sanctions, their impact may occur through different routes, and their contribution to future group-level cooperation may also be of different nature. We think that angry people punish because they perceive the unequal distribution to be morally unjust. Anger then induces retribution "whatever the consequences". This may also serve to deter future transgressions against all group members, including the punisher (e.g., Carlsmith, Darley, & Robinson, 2002). Guilt feelings, we think, are elicited when people perceive themselves in some way responsible for meting out punishment. When people anticipate guilt they punish in order to restore a sense of justice among group members in general and the victim in particular (e.g., Darley & Pitman, 2003). This ensures that we sometimes punish even if deterrence of future offenses is absent.

We thus expect that anger and guilt feelings are independent determinants of third-party sanctions. In two experimental studies, we tested our predictions by independently manipulating the elicitation of anger and guilt. Our data revealed that the elicitation of either emotion already resulted in punishment, as did the elicitation of both.

2 Experiment 1

We investigated the impact of anger and guilt feelings in a third-party punishment paradigm (Fehr & Fischbacher, 2004). In this paradigm, participants witness an unfair distribution of some valuable endowment between two other people. As a form of costly punishment, participants then had the opportunity to assign reduction-points to the allocator, out of their own endowment.

To independently manipulate the elicitation of anger and guilt, we varied the extent to which the unfair distribution *was made intentionally* by the allocator (as a proxy for anger) and the *participant's responsibility* for punishing the norm-violation (as a proxy for guilt). Research on ultimatum bargaining (e.g., Blount, 1995) shows that unfair proposals evoke angry reactions (resulting in increased rejection) only in case the outcome is an intentional act by another person, but not when people believe that, for instance, a computer has randomly (i.e., unintentionally) generated the offer. Similarly, the elicitation of guilt is related to perceptions of personal responsibility (e.g., Smith & Ellsworth, 1985). The bystander effect (Latané & Darley, 1970) shows that the presence of others diffuses personal responsibility and thereby decreases the willingness to help a victim. Simply put, this means that, as the number of people that are able to alleviate someone's misfortune increases, each individual feels less guilty for not helping and is therefore less willing to do so. Bystander effects have also been demonstrated for norm-enforcing behavior (Chekroun & Brauer, 2002).

Based on these well-documented effects, we predicted that third-party punishment would be increased (1) if norm-violations were brought about intentionally, and (2) if participants were solely responsible for norm-enforcement.

2.1 Method

Participants and design. Thirty-eight male and 53 female undergraduate students (M_{age} = 20.84 years) participated in exchange for course credit. The experimental design included two between-subject factors: Intentionality of norm-violation (intentional vs. unintentional) and Responsibility for norm-enforcement (high vs. low).

Procedure. Participants were seated in individual cubicles. Instructions and measures were presented on a computer. Participants were told they would participate in a single-shot negotiation, and would be randomly assigned to one of the three possible roles (see Appendix A). In

fact, all participants were assigned the role of third-party (denoted as Player C) and received an initial endowment of 50 points. Each point was worth one lottery ticket. Participants were told that they would first witness a distribution of 100 points between two other players, and that they would subsequently have the opportunity to adjust that distribution by assigning reduction points to the allocator (denoted as Player A). Each reduction point would reduce the endowment of the player to whom it was assigned by three points. Next, the negotiation started by the allocator dividing an initial endowment of 100 points with the receiver. All participants learned that the allocator kept 80 point and that the receiver received only 20 points. Hence, the allocator clearly violated the equal distribution norm.

Manipulation of intentionality. In the unintentional norm-violation condition participants learned that the proposal from the allocator was in fact generated randomly by the computer. In order to avoid confounding of both manipulations, it was stressed to our participants that the receiver was unaware that the offer was generated randomly by the computer but still believed that another participant had made the offer. Hence, participants' responsibility to step up for the receiver by punishing the norm-violation was unaffected by the intentionality manipulation.

Manipulation of responsibility. In order to manipulate responsibility for norm-enforcement, participants were either told they were the only Player C in the interaction (high-responsibility condition), or that there were two other participants who were also assigned this role (low-responsibility condition).

Measures. The amount of reduction-points (0–50) assigned to the allocator served as a measure of third-party punishment. To check the intentionality manipulation, participants indicated the extent to which they perceived *the allocator* as "acting intentionally", "responsible", and "accountable" (α = .59) *for the proposed distribution*. To check the responsibility manipulation, participants rated the extent to which they *themselves* felt "responsible", "accountable", and "liable" (α = .84) *for assigning reduction points*. All measures were assessed on 7-point scales from 1 (*not at all*) to 7 (*completely*).[1]

Pretest. In addition, we also tested whether our manipulations of intentionality and responsibility affected feelings of anger and guilt as we intended. However, because previous studies (e.g., Keltner, Locke, & Audrain, 1993) have shown that explicitly rating emotional reactions may attenuate their effect, we did not measure emotions in this study. Instead, we performed an independent check to verify whether our manipulations had the intended effects on participants' emotions. Hereto, we presented a separate sample of participants (N = 90) with four different hypothetical scenarios depicting the experimental conditions of the present study. Participants were then asked to indicate the extent to which they expected that observing the unequal distribution would make them feel angry or guilty over not assigning reduction points. Anger towards the allocator was assessed by the items "angry", "mad", and "annoyed" (α = .94). Anticipated guilt over not punishing was assessed by the items "guilty", "feeling bad for what I did", and "ashamed" (α = .85).[2] Ratings were made on scales from 1 (*not at all*) to 7 (*extremely*).

2.2 Results

Pretest. We first analyzed the data from the separate sample used to pretest the effect of our manipulations on emotions. A 2 (intentionality) $\times$ 2 (responsibility) ANOVA on anger showed only a main effect of intentionality, $F(1, 86) = 32.34$, $p < .001$, $\eta^2 = .27$. As intended, unequal offers by another participant aroused more anger (M = 3.29, SD = 1.40) than unequal offers made by a computer (M = 1.71, SD = 1.22). We also found only the expected main effect of responsibility on anticipated guilt $F(1, 86) = 28.60$, $p < .001$, $\eta^2 = .24$, showing that participants felt less guilty when there were other punishers (M = 2.68, SD = 1.31) than when they were the only person responsible for punishing the unfair distribution (M = 4.26, SD = 1.52).

Manipulation checks. Subsequently, we turned to the data from the sample used in the main study. A 2 (intentionality) $\times$ 2 (responsibility) ANOVA on the perceived intentionality of the allocator revealed that our manipulation was successful. We found only the intended main effect of intentionality, F(1,91) = 7.12, p = .005, one-tailed, η2 = .08. Participants in the computer-generated offer condition found that the allocator had acted less intentionally (M = 4.64, SD = 1.01) than participants in the person-generated offer condition (M = 5.25, SD = 1.14). The same ANOVA on perceived personal responsibility to punish yielded only a main effect of responsibility, F(1,91) = 11.53, p < .001, η2 = .57. Participants in the low-responsibility (i.e., multiple punisher) condition felt less responsible (M = 3.06, SD = 0.88) than in the high-

[1]Note, that the original text was in Dutch and that we used the following Dutch words (English in parentheses): moedwillig (intentionally), verantwoordelijk (responsible), aansprakelijk (accountable), and beslissend (liable).

[2]Note, that the original text was in Dutch and that we used the following Dutch words (English in parentheses):

boos (angry), kwaad (mad), geirriteerd (annoyed), schuldig (guilty), voelde me slecht over wat ik gedaan had (feeling bad for what I did), and schaamte (ashamed). Note moreover, that we assessed guilt by means of three items, including on item tapping feeling of shame. We realize that shame is an emotion distinct from guilt, but at the same time, both are self-conscious emotions that are felt when people feel responsible for something bad happens to another person. In the present study, this is also the case, as is evident from the high reliability of the scale.

responsibility (single-punisher) condition (M = 5.48, I = 1.30). We wish to stress that perceived responsibility was not affected by whether the offer was made by a computer or a person. Hence, manipulating intentionality had not affected participants' responsibility towards the receiver.

Punishment. A 2 (intentionality) × 2 (responsibility) ANOVA on the number of reduction points yielded a main effect of responsibility, F(1,89) = 7.61, p = .004, one tailed, $\eta 2$ = .08, indicating that participants punished more if they were the only punisher (M = 14.16, SD = 5.84), than when there were two other punishers present (M = 10.56, SD = 6.42). The hypothesized effect of intentionality was also (barely) significant (M = 13.42, SD = 6.78, when the offer was intentional; M = 11.13, SD = 5.82, when unintentional; F(1,89) = 2.90, p = .046, one-tailed, $\eta 2$ = .03).[3]

2.3 Discussion

In this experiment we manipulated the extent to which norm violations were brought about intentionality (as a proxy for anger) and people's responsibility for sanctioning that norm violation (as a proxy for guilt). A pretest demonstrated that the manipulations of intentionality and responsibility affected feelings of anger and guilt as intended. This results therefore support the notion that anger and guilt both elicit third-party sanctions of approximately 27 percent of the endowment. Levels of punishment were reduced to about 16% of their endowment only if people were not responsible for punishing an unintentional norm violation. Because the actual punishment was three times the number of reduction points, the punishments were quite substantial.

In order to further test the independent contribution of anger and guilt on punishment, we ran another experiment in which we investigated the effects of a manipulation that independently *inhibited* anger and guilt rather than *eliciting* each emotion separately. The second experiment provided two other important extensions. First, levels of anger and guilt were not directly associated to levels of punishment in Study 1 and therefore, their impact could only be inferred. Although we had good reasons to initially refrain from direct assessment of emotional reactions in the first experiment, the independent causal influence of anger and guilt on third-party sanctions still requires empirical demonstration, which we did in Experiment 2.

Secondly, participants' willingness to punish a computer (in the unintentional violation condition of Experiment 1) may seem surprising at first. In our opinion, this finding clearly shows that people besides seeking retaliation, also punish norm-violations to restore a sense of justice in the harmed person, particularly if this is the only means to alleviate the victim's suffering. We acknowledge however, that participants in our study may have considered punishing a computer an awkward way to express their concern with the victim. Another reason for conducting Experiment 2 was to alleviate doubts about the validity of this effect, by using a different manipulation of anger that did not require participants to punish a computer.

3 Experiment 2

To independently manipulate the elicitation of anger and guilt without requiring our participants to punish an inanimate object, we applied a noise-manipulation (Van Lange, Ouwerkerk & Tazelaar, 2002). In social interactions, noise can be defined as any kind of involuntary disturbance from an intended outcome. Positive noise causes the actual outcome to be better than originally intended and negative noise leads to worse outcomes.

In the present study, we introduced noise by stating that there was a possibility that the computer would randomly change the allocator's offer to the receiver, without the receiver being aware of this possibility. In the positive noise condition, an unfair division by the allocator was increased to a more or less equal offer. Notably, participants believed the receiver to be unaware of the original proposal and also of the possibility that the original proposal could be randomly changed. So, feelings of anticipated guilt should not be affected in the positive noise condition. Hence, positive noise reduced the need to restore justice to the receiver, but did not affect the deterrence function of third-party sanctions. Consequently, participants should feel angry towards the allocator for making an unequal offer.

Negative noise was modeled by the reduction of a fair offer to an unfair one. This manipulation maintained the need to restore a sense of justice to the receiver, but not to deter future norm violations by the allocator. We expected this to evoke anticipated guilt for not punishing but to inhibit anger towards the allocator.

Both conditions were compared to a control condition in which the allocator made an unfair offer that was not changed. We expected this to elicit both feelings of anger and anticipated guilt. Consequently we anticipated lower levels of punishment in both noise-conditions, showing the unique contribution of anger to punishment in the pos-

[3] We found a nearly significant interaction, F(1,91) = 3.04, p = .085, two-tailed, $\eta 2$ = .03. Simple-effects analysis revealed that the main effect of responsibility on punishment was significant only within the unintentional violation condition, F(1,91) = 10.29, p = .002, $\eta 2$ = .11. Intentional norm violations were punished equally irrespective of the level of responsibility of the participants, F(1,91) = 0.53, p = .47, $\eta 2$ = .01. The interaction term is not included in the main analysis just reported; if it is included, the effect of intentionality is not quite significant (p=.054, one-tailed).

itive noise condition and of anticipated guilt in the negative noise condition.

Furthermore, we predicted that the effects of noise would be mediated by anger and anticipated guilt. Specifically, the difference in observed levels of punishment between the control and the positive noise condition should be mediated by feelings of anticipated guilt. Feelings of anger should be equally high in the control and the positive noise condition, which are similar in the sense that a reaction is required to an unfair offer by the allocator. The difference in observed levels of punishment between the control and the negative noise condition on the other hand, should be mediated by feelings of anger. Feelings of anticipated guilt should both be high in the control and the negative noise condition, which are similar in the sense that a reaction is required to an unfair outcome to the receiver. Hence, we predict the effects of positive and negative noise to be mediated by the feelings that they are intended to reduce, not by the feelings that they should not affect compared to the control condition.

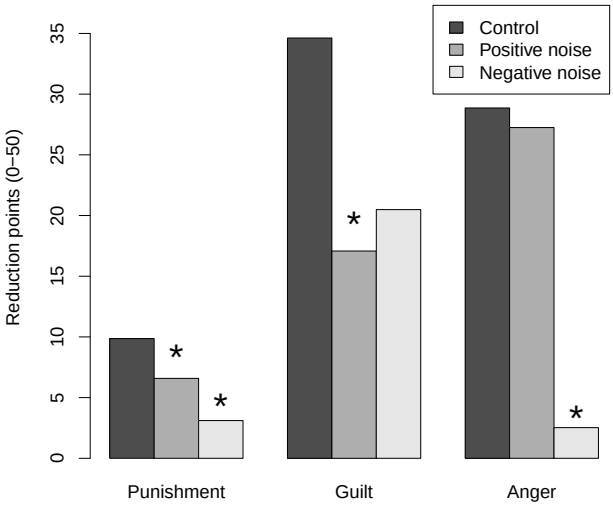

Figure 1: Mean number of reduction points assigned in the control, positive, and negative noise condition in Experiment 2. * denotes a significant difference in means from the Control condition at $p < .05$.

3.1 Method

Participants and design. Twenty-seven male and 103 female undergraduate students (M_{age} = 19.3 years) participated in exchange for course credit. Procedures and instructions were identical to Experiment 1, except for the noise-manipulation (see Appendix B), which introduced the following three conditions: Control (no change of unfair offer), Positive noise (unfair offer increased), and Negative noise (fair offer decreased).

Noise manipulation. As stated, participants were instructed that in some instances, the computer would randomly change the offer made by the allocator (Van Lange et al., 2002). This change affected only the outcome to the receiver, whereas the allocator would still receive the payoff as originally proposed. In the no-change control condition, the receiver received 20 (out of 100) points from the allocator. In the positive noise condition, the allocator made the same unequal (i.e., 80/20) offer, yet the computer increased this offer to 52 points for the receiver. In the negative noise condition, the computer reduced an initially fair (i.e., 50/50) proposal to a mere 18 points for the receiver. We chose just off-round figures in the noise conditions to render the ostensibly random nature of changes more credible to participants.

Measures. After participants learned the offer (and how this was changed in the noise conditions) they indicated the extent to which they felt angry (α = .96) and guilty (α = .91), using the same items as in the pretest of Experiment 1. Ratings were made by dragging a pointer on a 100 point visual analogue scale, anchored *not at all — extremely*. Next, they indicated the number of reduction-points (0–50) assigned to the allocator as a measure of third-party punishment.

3.2 Results

Figure 1 shows the main results. As predicted, participants in the control condition (M = 9.86, SD = 7.19) punished more than participants in the positive noise condition (M = 6.59, SD = 5.75), $t(82) = -2.30$, $p = .012$, one-tailed. Participants in the control condition also punished more than participants in the negative noise condition (M = 3.11, SD = 5.68), $t(87) = -4.93$, $p < .001$.[4]

Moreover, we also found the expected differences between conditions in terms of self-reported anger towards the allocator, and guilt over not assigning reduction points. The control (M = 34.62, SD = 31.10) and the positive noise condition (M = 17.07, SD = 27.96) differed significantly only in terms of anticipated guilt over not punishing (p = .004, one-tailed) but not in terms of anger towards the allocator (p = .396). The control and the negative noise condition (M = 20.49, SD = 24.48) differed in terms of anticipated guilt (p = .010), but the control (M = 28.86, SD = 26.79) condition differed much more from the negative noise condition (M = 2.53, SD = 10.61) in terms of anger towards the allocator ($p < .001$). A test of the canonical correlation between anger and guilt, on the one hand, and positive/control contrast and negative/control contrast, on the other, revealed that the second canonical correlate was significant (p=.004, using Rao's approximation; see Burns, 2009), which indicates

[4]The difference between the positive and the negative noise condition was also significant, $t(85) = -2.83$, $p = .01$, two-tailed.

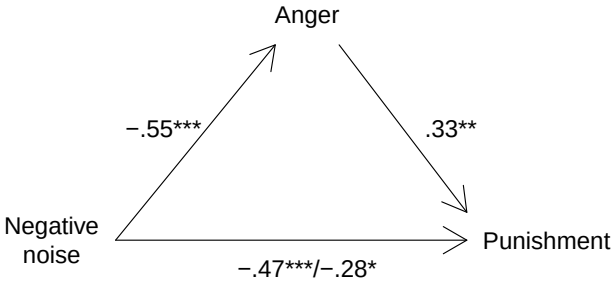

Figure 2: Influence of negative noise, direct and mediated by anger on punishment in Experiment 2. Numbers represent standardized regression coefficients: * $p < .05$. ** $p < .01$. *** $p < .001$.

that two distinct factors were required to predict emotion from experimental condition.

The effects of negative noise on punishment were mediated by anger, as we hypothesized.[5] Specifically, as shown in Figure 2, in regressions using standardized variables, negative noise was associated with reduced anger ($\beta = -.47$, p < .001) and with reduced punishment ($\beta = -.55$, p < .001). And anger predicted punishment when negative noise was included in the model ($\beta = .33$, p < .001). The effect of negative noise on punishment was reduced from $-.47$ to $-.26$ when anger was included in the model.

However, there was no mediation of the effect of positive noise by guilt feelings. Positive noise did reduce punishment ($\beta = -.25$, p=.024), and it reduced guilt ($\beta = -.28$, p=.008), but guilt did not affect punishment significantly when positive noise was included in the model ($\beta = .06$, p=.614).[6]

3.3 Discussion

In this experiment we aimed to inhibit feelings of anger and guilt independent from each other. To that end we introduced a noise manipulation (Van Lange et al., 2002) that entailed either an increase of an unfair offer to an equal one (positive noise) or a decrease of a fair offer to an unequal one (negative noise). Consequently, the positive noise manipulation reduced the responsibility of the participant for punishing a norm violation that was nevertheless still intentional. The negative noise manipulation maintained the participant's responsibility to punish an unequal offer that was however not made intentionally. We compared the effects of reduced intentionality and responsibility on punishment levels to those in a control condition in which an unfair offer was made intentionally and participants were fully responsible for punishing this norm violation. As predicted, we found that reducing intentionality and responsibility each resulted in a significant reduction of punishment levels. These findings complemented those of Experiment 1 in which an increase rather than a decrease of intentionality and responsibility was also found to exert an independent effect on third party punishment.

Although a canonical correlation analysis suggested that the effects of our noise manipulations were best predicted by two distinct emotion factors, we found direct confirmation only of the hypothesized role of anger. Results supported anger as a determinant of third-party punishment in the negative noise condition. Anger also apparently mediated the effect of negative noise on punishment. But guilt feelings did not mediate punishment in the positive noise condition, contrary to our hypothesis.

The failure of guilt to affect punishment directly does not necessarily contradict its role. The feeling of guilt might have been reduced by the intention to punish. Thus, guilt could drive the desire to punish but then be reduced once this desire is translated into an intention to act. Such an interpretation is consistent with the effects of positive noise on guilt, and on punishment. Anger, by contrast, would not dissipate so quickly, perhaps not until the punishment is actually implemented.

4 General discussion

We found in two experiments that third-party punishment has distinct determinants, that probably relate to distinct emotional processes. Experiment 1 revealed that manipulating the intentionality of the norm-violation (as a proxy for anger) and the responsibility for sanctioning (as a proxy for guilt) both increase levels of punishment. Experiment 2 demonstrated the involvement of anger and, possibly, anticipated guilt more directly, by showing that inhibiting each emotion reduced punishment. Together, these findings suggest that both emotions separately constitute sufficient but not necessary causes of third-party punishment. Apparently, punishment can be extended by third parties in response to unfair intentions and to unfair outcomes. When either is missing, punishment is reduced.

These results both support and extend current proposals about the functions of sanctioning norm-violations. The impact of anger is in line with views that hold punishment to primarily serve retaliatory purposes (Carlsmith et al. 2002; Darley & Pittman, 2003). Additionally, we have provided initial support that feelings of anticipated guilt may underlie third-party sanctions as well. Feelings

[5] See Baron and Kenney (1986) and MacKinnon et al., (2002) for discussion of mediation.

[6] In addition, anger did not mediate the effects of positive noise and guilt did not mediate the effects of negative noise. These were not hypothesized to occur.

of guilt over not punishing norm violations may themselves suffice to cause punishment. That is, anticipated guilt instigates punishment even when people know that the norm was not (intentionally) violated. Identifying anticipated guilt as an additional determinant broadens the functional perspective of third-party sanctions. Feelings of guilt generally motivate behavior that is aimed to restore transgressions towards others (Haidt, 2003), guilt-induced punishment may imply that punishers seek to compensate the victim by restoring a sense of justice (e.g., Darley & Pitman, 2003) rather than by retaliating against the perpetrator of a violated norm.

It may seem that our findings contradict those of a recent study that also investigated the consequences of a mismatch between intended and actually obtained outcomes in a dictator game (Cushman, Dreber, Wang, & Costa, 2009). Whereas our findings seem to imply that intentions matter more than outcomes, Cushman and colleagues (2009) observed that unfair outcomes have a strong effect on the level of punishment, even if allocators apparently had fair intentions. Two important differences with our own study are that Cushman et al. investigated second-party sanctions (i.e., by the victim rather than a third party), and that their manipulation allowed for idiosyncratic attribution of intentionality by the victim. This resulted in the victim having ample leeway to ascribe unfair intentions to the allocator in case of an unequal distribution, which may have accounted for the finding that outcomes were the primary cause of punishment in their study. Our manipulation of noise in Study 2 guaranteed a stricter differentiation of the effect of intention and outcome on punishment levels, which we believe explains the fact that outcome is not the strongest determinant of punishment in our research.

Our results should be interpreted bearing some reservations in mind. One particular aspect of our studies may be of concern to perceptive readers. As participants' outcomes in our studies earned them lottery tickets, and their chances of winning the lottery depended upon the number of tickets they earned themselves but also on the number of tickets earned by the other participants, one may argue that they punished in order to increase their chances of winning the lottery. We consider this unlikely however, as the levels of punishment in our studies did not exceed those reported in other third-party punishment experiments in which punishment actually cost money and therefore could in no way improve punishers' outcomes (e.g., Fehr & Fischbacher, 2004). We therefore do not believe that participants' motivation to win the lottery has confounded the present findings.

A similar line of reasoning may cause one to wonder if our experiments induced only anger and guilt. For instance, in Experiment 1, when participants learn that they receive 50 points while Player A has 80 points, feelings of envy could be evoked. Envy is the emotion that motivates behavior aimed at reducing the differences between oneself and another who is better off (e.g., Van de Ven, Zeelenberg, Pieters, 2009). Hence participants may be willing to incur a cost to punish the other, and hence reduce inequality. We agree that such effects of envy could exist, also in our studies, in addition to the effects we have found. Future studies could investigate the extent to which envy has an additional effect here.

Furthermore, we documented independent contributions of anger and guilt only in a single paradigm. Whether anger and guilt underlie third-party sanctions in general remains to be seen. Similarly, we studied punishment only in response to violations of a single norm that prescribes equal distribution. Behavior in other situations may be guided by different norms (e.g., courage, loyalty, and modesty). Whether or not the violation of other norms elicits feelings of anger and guilt, as well, also remains to be seen. Some studies seem to suggest that different types of violations evoke specific emotional reactions (Rozin, Lowery, Imada, & Haidt, 1999). This would suggest that the emotional basis for punishing norm violations is more diverse than the present study suggests. Documenting specific emotional reactions may prove a fruitful way to establish functional links between different types of norm violations and the various goals underlying their punishment (Carlsmith et al., 2002; Darley & Pitman, 2003; Zeelenberg, Nelissen, Breugelmans, & Pieters, 2008).

Insight about multiple motives underlying sanctions may also further our understanding of group-identification effects that have been reported for the punishment of norm violations. For instance, punishment is more severe when in-group members suffer from norm-violations than when out-group members are victimized (Bernard, Fischbacher & Fehr, 2006). On the other hand, more severe punishment befalls in-group members violating social-norms than when out-group members harm each other (Shinada, Yamagishi & Ohmura, 2004).

It is likely that the motives underlying specific instances of punishment depend upon the relations between punisher, perpetrator, and victim. Closer ties with the perpetrator, for instance, may attenuate the elicitation of anger as punishers may be more inclined to justify violations from people with whom they have a close relationship. Guilt feelings are apt to be responsive to affiliations with the victim as they are likely to affect punishers' perceptions of responsibility. Alternatively, it has been found that increased identification with the victim enhances feelings of anger (Yzerbyt, Dumont, Wigboldus & Gordijn, 2003), which may imply that closer connections with either party would instigate more intense feelings of anger in punishers. Although Yzerbyt and colleagues did not assess feelings of anticipated guilt, it may

be that feelings of guilt for not punishing norm-violations are not dependent upon ties with either victim or perpetrator. Rather, guilt-induced punishment in order to alleviate a victim's suffering may reflect the violation of personal standards of fairness.

The compound causation of third-party sanctions deserves further inquiry as it may improve policies aiming to stimulate informal (i.e., non-institutionalized) sanctions as a means to instill norms that benefit everyday social interactions. Whereas the execution of actual third-party punishment is arguably the function of governments, and actual third party punishment is rare in everyday interactions (Sabini & Silver, 1982), people will not refrain from extending mild forms of punishment in terms of remarks of passive sanctions, as the familiar examples in the opening paragraph illustrated. For instance, with respect to phenomena like random violence, littering, and proper codes of conduct in particular environments, informal sanctions can improve social interactions. Regarding such issues, it should be noted that punishment is still sensitive to concerns of self-interest, for increasing costs of sanctions appear to reduce the level of punishment (Fehr & Fischbacher, 2004). Considering multiple motives may help to identify factors that maintain informal sanctions even under conditions that are apt to undercut other incentives. Angry reactions may dampen when norm-violations also benefit the third-party, for example. Still, it may well be that guilt-induced punishment prevails even under conditions where punishers do not feel angry.

To summarize, the present study is the first to empirically support the proposed link between anger and third-party punishment. Moreover, we showed that angry reactions to norm-violations are not necessarily required and third-party punishment may also be motivated by previously unexplored emotions like guilt. Either emotional response may be in itself sufficient to induce punishment. This suggests that we not only punish norm violations to deter future transgression by the perpetrator, but also to restore a sense of justice to the victim.

References

Barron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and social Psychology, 51,* 1173–1182.

Batson, C. D. (2006). Not all self-interest after all: Economics of empathy-induced altruism. In D. De Cremer, M. Zeelenberg, & J. K. Murnighan (Eds), *Social psychology and economics* (pp. 281–299). Mahwah, NJ: Lawrence Erlbaum Associates.

Batson, C. D., Kennedy, C. L., Nord, L., Stocks, E. L., Fleming, D. A., Marzette, C. M., et al. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology, 37,* 1272–1285.

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., & van Winden, F. (2007). Reciprocity and emotions in bargaining using physiological and self-report measures. *Journal of Economic Psychology, 28,* 314–323.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442,* 912–915.

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes, 63,* 131–144.

Bosman, R., & van Winden, F. (2002). Emotional hazard in a power to take experiment. *The Economic Journal, 112,* 147–169.

Burns, C. T. (2009). *yacca: Yet another canonical correlation analysis package*, (R package version 1.1). http://CRAN.R-project.org.

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.

Checkroun, P., & Brauer, M. (2002). The bystander effect and social control behavior: The effect of the presence of others on people's reactions to norm violations. *European Journal of Social Psychology, 32,* 853–867.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "Trembling Hand" game. *Plos One, 4,* 1–7.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review, 7,* 324–336.

De Hooge, I. E., Breugelmans, S. M. & Zeelenberg, M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion, 21*, 1025–1042.

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

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

Frank, R. (2004). Introducing moral emotions into models of rational choice. In A. S. R. Manstead, N. Frijda & A. Fischer (Eds.), *Feelings and emotions: The Amsterdam symposium* (pp. 422–440). Cambridge: Cambridge University Press.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). NY: Oxford University Press.

Hume, D. (1739/1969). *A treatise of human nature.* London: Penguin.

Keltner, D., Locke, K. D., & Audrain, P. (1993). The influence of attributions on the relevance of negative feelings to personal satisfaction. *Personality and Social Psychology Bulletin, 19,* 21–29.

Ketelaar, T., & Au, W. T. (2003). The effects of feelings of guilt on the behavior of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion, 17,* 429–453.

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* NY: Appleton-Century-Crofts.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104.

Nelissen, R. M. A., Dijker, A. J., & de Vries, N. K. (2007). How to turn a hawk into a dove and vice versa: Interactions between emotions and goals in a give-some dilemma game. *Journal of Experimental Social Psychology, 43,* 280–286.

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes, 68,* 208–224.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and social Psychology, 76,* 574–586.

Sabini, J., & Silver, M. (1982). *Moralities of Everyday Life*. New York: Oxford University Press.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300,* 1755–58.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies: "Altruistic" punishment of in-group members. *Evolution and Human Behavior, 25,* 379–393.

Smith, A. (1759/1976). *The theory of moral sentiments.* Oxford: Clarendon Press.

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48,* 813–838.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46,* 35–57.

Tooby, J., & Cosmides, L. (1990). The Past Explains the Present: Emotional Adaptations and the Structure of Ancestral Environments. *Ethology and Sociobiology, 11*, 375–424.

Ven, N. van de, Zeelenberg, M., & Pieters, R. (2009). Leveling up and down: The experiences of benign and malicious envy. *Emotion*, *9*, 419–429.

Van Lange, P. A. M., Ouwerkerk, J. W., & Tazelaar, M. J. A. (2002). How to overcome the detrimental effects of noise in social interaction: The benefits of generosity. *Journal of Personality and Social Psychology, 82,* 768–780.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*, 110–116

Yamagishi, T. (1988). Seriousness of social dilemmas and the provision of a sanctioning system. *Social Psychology, 51*, 32–42.

Yzerbyt, V., Dumont, M., Wigboldus, D., & Gordijn, E. (2003). I feel for us: The impact of categorization and identification on emotions and action tendencies. *British Journal of Social Psychology, 42,* 533–549.

Zeelenberg, M., Nelissen, R. M. A., Breugelmans, S. M., & Pieters, R. (2008). On emotion specificity in decision making: Why feeling is for doing. *Judgment and Decision making*, *3*, 18–27.

Appendix A: Instructions for Experiment 1

Dear participant,

Welcome to this study on social interactions. Please read the following instructions carefully. In case of questions or uncertainties, please call the experimenter by pushing the red button on the intercom.

You are going to participate in an interaction between three parties. All participants will be randomly assigned to either one of three possible roles, Player A, B, or C.

In the course of this interaction, you and the other players can earn points. For each point that you have earned you will get one lottery ticket. At the end of this week, when the experiment is over, we will draw three tickets that will win a price of €50,- each. This means that your decisions in this interaction may increase or decrease your chances of winning money.

Before we start, the computer will now randomly assign each participant to the role of either Player A, B, or C.

[TIME LAG]

Conditions:

Single Punisher: You are assigned to the role of Player C.

Multiple Punishers: You are assigned to the role of Player C. Apart from yourself; two other participants have also been assigned the role of Player C.

Player A = Participant: In this interaction you are paired with two other participants who have been assigned the roles of Player A and Player B.

Player A = Computer: The role of Player A in this interaction will be played by the computer. The computer will randomly allocate points to Player B. Player B, however, is not aware of this and assumes that another participant makes the decision to allocate points.

Everyone will remain completely anonymous during and after the experiment. So, you will never know which other participants were the Players A and B in this interaction, nor will they know that you were the Player C with whom they had interacted.

If you press "Continue" you will receive information about the two phases of the social interaction. Players A and B will receive the same information.

In phase 1 only the participants that have been assigned to the role of Player A will make a decision. At the beginning of phase 1, Player A has 100 points. Player B has no points, and Player C has 50 points.

Player A may allocate a voluntary number of points to Player B. It is up to Player A to decide how many points he or she will allocate to Player B. Player A may even decide not to allocate any points at all to Player B, or to allocate all points to Player B. [EXAMPLES].

In phase 2 only the participants that have been assigned the role of Player C will make a decision. As soon as Player A has decided how many points to allocate to Player B, Players B and C will be informed of this decision.

Next, Player C can assign reduction points to Player A. Every reduction point that Player C assigns to Player A reduces the total number of points from Player C by 1, but the total number of points from Player A by 3. Player C is free to assign any number of reduction points to Player A, either zero of all 50 points that were initially assigned. [EXAMPLES].

So, Player A will first decide how to distribute the 100 points. As soon as Player A has made a decision, Players B and C will be informed thereof. Next, Player C may decide to assign reduction points to Player A.

Finally, player B will be informed of how much reduction points Player C has assigned to Player A.

Please press "Continue" to start with the interaction.

Appendix B: Instructions for Experiment 2

Dear participant,

Welcome to this study on social interactions. Please read the following instructions carefully. In case of questions or uncertainties, please call the experimenter by pushing the red button on the intercom.

You are going to participate in an interaction between three parties. All participants will be randomly assigned to either one of three possible roles, Player A, B, or C.

In the course of this interaction, you and the other players can earn points. For each point that you have earned you will get one lottery ticket. At the end of this week, when the experiment is over, we will draw three tickets that will win a price of €50,- each. This means that your decisions in this interaction may increase or decrease your chances of winning money.

Before we start, the computer will now randomly assign each participant to the role of either Player A, B, or C.

[TIME LAG]

You are assigned to the role of Player C.

Everyone will remain completely anonymous during and after the experiment. So, you will never know which other participants were the Players A and B in this interaction, nor will they know that you were the Player C with whom they had interacted.

If you press "Continue" you will receive information about the two phases of the social interaction.

In phase 1 only the participants that have been assigned to the role of Player A will make a decision. At the beginning of phase 1, Player A has 100 points. Player B has no points, and Player C has 50 points.

Player A may allocate a voluntary number of points to Player B. It is up to Player A to decide how many points he or she will allocate to Player B. Player A may even decide not to allocate any points at all to Player B, or to allocate all points to Player B. [EXAMPLES].

In phase 2 only the participants that have been assigned the role of Player C will make a decision. As soon as Player A has decided how many points to allocate to Player B, Players B and C will be informed of this decision.

Next, Player C can assign reduction points to Player A. Every reduction point that Player C assigns to Player A reduces the total number of points from Player C by 1, but the total number of points from Player A by 3. Player C is free to assign any number of reduction points to Player A, either zero of all 50 points that were initially assigned.

[EXAMPLES].

So, Player A will first decide how to distribute the 100 points. As soon as Player A has made a decision, Players B and C will be informed thereof. Next, Player C may decide to assign reduction points to Player A.

Finally, player B will be informed of how much reduction points Player C has assigned to Player A.

To Player C: Please not that in some instances, the computer will change the number of points allocated by Player A to Player B. The computer will randomly subtract or add points to the number of points allocated to Player B. It may also be the case that the number of points

will not be changed by the computer.

It is important to know that neither Player A nor Player B is aware of the possibility that the number of points may be changed by the computer. So, Player B will always assume that the number of points that he or she receives is the number that was assigned by Player A.

Please press "Continue" to start with the interaction.