The unresponsive avenger:

More evidence that disinterested third parties do not punish altruistically

Eric J. Pedersen[1, 2], William H. B. McAuliffe[1], & Michael E. McCullough[1]

[1]Department of Psychology, University of Miami

[2]Department of Psychology and Neuroscience, University of Colorado Boulder

Article in press at *Journal of Experimental Psychology: General*

Author Note

This work was supported by grants from the Air Force Office of Scientific Research (FA9550-12-1-0179) to M.E.M., the Arsht Research on Ethics and Community Program at the University of Miami to M.E.M, the John Templeton Foundation (29615) to M.E.M, an NSF Graduate Research Fellowship to E.J.P, and the Expanding the Science and Practice of Gratitude Project run by UC Berkeley's Greater Good Science Center in partnership with UC Davis with funding from the John Templeton Foundation. Some of the results in this paper were presented at the annual meeting of the Society for Personality and Social Psychology in 2016 and 2017.

Correspondence should be addressed to Michael E. McCullough (mikem@miami.edu), Department of Psychology, University of Miami, P.O. Box 248185, Coral Gables, FL 33124-0751, or Eric J. Pedersen (eric.j.pedersen@colorado.edu), Department of Psychology and Neuroscience, University of Colorado Boulder, 345 UCB Boulder, CO 80309-0345.

Abstract

Many social scientists believe humans possess an evolved motivation to punish violations of norms—including norm violations that do not harm them directly. However, most empirical evidence for so-called altruistic punishment comes from experimental economics games that create experimental demand for third-party punishment, raising the possibility that the third-party punishment uncovered in these experiments has been motivated by a desire to *appear concerned about social norms* rather than by actual concern about upholding them. Here we present the results of five experiments in which we used an aggression paradigm to contrast second-party and third-party punishment with minimal experimental demand. We also summarize the results of these experiments meta-analytically. We found robust evidence that subjects who were insulted by a stranger experienced anger and punished the insulter. To a lesser degree, subjects who witnessed a friend receive an insult also became angry and punished the insulter. In contrast, we found robust evidence that subjects who witnessed a stranger receive an insult *did not* punish the insulter, although they did experience modest amounts of anger. In only one experiment did we find any punishment on behalf of a stranger, and this result could plausibly be explained by the desire to escape the moral censure of other bystanders. Our results suggest that experimental designs that rely on demand-laden methods to test hypotheses about third-party punishment may have overstated the case for the existence of this trait.

Keywords: punishment, cooperation, anger, empathy, friendship, welfare interdependence

Introduction

For the past 25 years, social scientists have been intensively studying how humans use punishment to enforce cooperation. These efforts have given rise to the proposal that humans possess an evolved moral motivation to engage in third-party punishment of norm-violators—that is, punishment of norm-violators who have not harmed the punisher directly. Humans' apparent penchant for third-party punishment is thought by many to play a key role in stabilizing cooperation among non-kin (Boyd & Richerson, 1992; Fehr & Fischbacher, 2003; Fehr & Gächter, 2002; Gintis, 2000).

In light of results from experimental economics games and theoretical models, many researchers have conceptualized third-party punishment as *biologically altruistic* because it is costly to the punisher and beneficial for the victim on whose behalf the punisher intervenes (e.g., Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Fischbacher, 2003, 2004; Fehr & Gächter, 2002; Henrich et al., 2005). On this view, third-party punishment results from an evolved propensity to punish interpersonal harm-doers at a personal cost—that is, *when there is no possibility for the punisher to receive direct benefits (or even indirect ones, such as the benefit that comes from helping a genetic relative) from doing so.* Most experiments on third-party punishment have focused on interactions among anonymous strangers, which on its face seems to rule out the possibility that putative punishers might expect to benefit from their action. For instance, punishers in an anonymous non-iterated game cannot recoup the fitness costs of punishing through reputational gains (Barclay, 2006; Jordan, Hoffman, Bloom, & Rand, 2016), a new reciprocal relationship with the victim of the stingy player's decision (Trivers, 1971), or the enhanced welfare of a genetic relative (Hamilton, 1964). Even so, the claim that humans possess a robust propensity for altruistic punishment has been challenged both theoretically (Baumard,

André, & Sperber, 2013; Burnham & Johnson, 2005; Hagen & Hammerstein, 2006; McCullough, Kurzban, & Tabak, 2013; West, El Mouden, & Gardner, 2011) and empirically (Krasnow, Cosmides, Pedersen, & Tooby, 2012; Pedersen, Kurzban, & McCullough, 2013).

The strongest empirical support for the altruistic punishment hypothesis comes from experiments that have used the third-party punishment game, which is a three-player extension of the dictator game (Fehr & Fischbacher, 2004; for earlier economic games with similar designs, see Kahneman, Knetsch, & Thaler, 1986; Turillo, Folger, Lavelle, Umphress, & Gee, 2002). This game involves a Dictator who chooses to give any portion of an endowment of money ($10, say) to a second player, the Receiver, who has no influence on the Dictator's decision. The third player is an Adjuster who is made aware of his or her ability to pay a cost to reduce the Dictator's earnings following the Dictator's decision. Importantly, the Adjuster's earnings are completely unaffected by the Dictator's behavior toward the Receiver: The only way the Adjuster's monetary outcome can be affected is if he or she decides to pay to punish the Dictator. Typically, Adjusters incur a personal cost to punish Dictators who fail to share "fair" amounts (i.e., less than 50%) of the endowment with the Receiver. For example, in the original experiment, roughly 60% of Adjusters punished unfair Dictators, and Adjusters' punishment expenditures were directly proportional to the Dictator's unfairness (Fehr & Fischbacher, 2004).

These results have been replicated in several cultures (Bernhard, Fischbacher, & Fehr, 2006; Henrich et al., 2005; Henrich et al., 2010; Henrich et al., 2006; Marlowe et al., 2008; cf. Marlowe, 2009) and obtain both when experimenters elicit Adjusters' punishment decisions *ex post* using the "game method" (i.e., when Adjusters make their punishment decisions only after they have learned of the Dictator's decision) and when they elicit them *ex ante* using Selten's (1967) "strategy method" (i.e., when Adjusters' punishment decisions are behavioral

commitments in advance of learning the Dictator's actual decision; Jordan, McAuliffe, & Rand, 2015). Self-report evidence from several studies also suggests that Adjusters' decisions to punish are regulated proximately by anger in response to violations of social norms (Fehr & Fischbacher, 2004; Fehr, Fischbacher, & Gächter, 2002; Fehr & Gächter, 2002; Herrmann, Thöni, & Gächter, 2008). This angry emotional response is called "moral outrage" in social psychology to signify that it is elicited not by harms incurred by the self or a cared-for other, but rather by violations of moral principles (O'Mara, Jackson, Batson, & Gaertner, 2011). Thus, third-party punishment is conceptualized as *altruistic* because its putative evolved function is to benefit others at the net cost to the self, and *moral* because it supposedly results from a perceived norm violation and subsequent moral outrage.

Altruistic third-party punishment: Altruism or artifact?

Although evidence from the third-party punishment game is consistent with a view of third-party punishment as altruistic, a few methodological features of the standard game might conspire to produce extensive experimental demand for punishment (Weber & Cook, 1972), which raises the question of whether the behavior the game elicits really does emerge from a desire to punish norm-violators. Two features stand out in particular. First, Adjusters are presented with only two behavioral options: to punish the Dictator or do nothing. They cannot reward the Dictator or compensate the Receiver, for instance. This limited behavioral range makes obvious to subjects that it is their propensity to punish that is of experimental interest. It also restricts subjects to a single behavioral option for fulfilling their desires (no matter what those desires might be). This is not an idle worry: Experiments that afford subjects with multiple options for action have revealed that (a) people often prefer to compensate victims of unfairness rather than punish transgressors, and (b) punishment is highest when there are no options for

compensation (Chavez & Bicchieri, 2013; Leliveld, Dijk, & Beest, 2012; Lotz, Okimoto, Schlösser, & Fetchenhauer, 2011; see also Balafoutas, Nikiforakis, & Rockenbach 2014, 2016 for related results in field studies[1]). Furthermore, parallel research on the Dictator Game itself, whose scores are often interpreted as measures of altruism or the desire to be fair, also show that the orthodox meanings ascribed to scores from experimental economics games that possess high levels of experimental demand can hardly be taken for granted. When Dictators' behavioral options are widened—for example, when they can not only share money with the Receiver but also to take money—generosity declines precipitously, in some cases to zero (for examples, see: Bardsley, 2008; List, 2007). The fact that such a small methodological change creates such drastic reductions in Dictators' offers implies that those offers are motivated by something other than a straightforward desire to be generous or fair.

Second, the third-party punishment game elicits behavior via explicit prompts. In the typical game, experimenters elicit punishment decisions from Adjusters by asking them whether (and if so, by how much) they would like to adjust the Dictator's monetary earnings either in light of *ex post* knowledge of how the Dictator treated the Receiver (when the Game Method is used), or as *ex ante* commitments conditional on the Dictator's yet-to-be-revealed treatment of the Receiver (when the Strategy Method is used). The experimenter's explicit request for behavior creates common knowledge between experimenter and subject (Thomas, De Freitas, DeScioli, & Pinker, 2016): By explicitly asking subjects to indicate their interest in spending

[1] Balafoutas, Nikiforakis, & Rockenbach (2014, 2016), as well as Balafoutas & Nikiforakis (2012) stand out in this literature as notable exceptions with regard to their lack of experimental demand for punishment. In their field studies, they report punishment rates of littering that range from 4% to 17%, which they interpret as evidence of altruistic punishment. However, an equally plausible interpretation is that punishers in these situations engage in punishment to deter littering in their own local environments (i.e., that the punishment they seem willing to administer is actually a form of second-party punishment). Thus, it is unclear exactly how to relate their findings to the literature on third-party punishment, though their ingenious use of field studies to study punishment in general can be fuitfully adopted to directly study third-party punishment in the future.

money to impose a retaliatory cost on a stingy Dictator, the subject realizes that the researcher (and possibly also the victim) knows that the subject knows that he or she now has the ability to take action against the Dictator in response to the Dictator's stingy behavior toward the Receiver.

Establishing common knowledge in this fashion is especially problematic for interpreting the meaning of scores in the third-party punishment game because unfair treatment and (proportionate) third-party sanctions in response to it are viewed by most people as morally bad and morally good, respectively (Baumard et al., 2013). Once common knowledge has been established around the Adjuster's possession of a means of sanctioning a stingy Dictator, Adjusters might become motivated to punish to signal their moral disapproval of norm violations and their moral approval of sanctions for norm-violators, *even if they actually possess little or no motivation to pay a personal cost in order to punish the norm-violator per se.* On this view, the creation of common knowledge elicits the desire *to appear* altruistically punitive, but not the desire *to actually be* altruistically punitive. In recent research on the Dictator Game that has explicitly explored this line of reasoning about the influence of common knowledge on prosocial behavior, the removal of features that established common knowledge between experimenters and Dictators reduced Dictators' mean monetary transfers to values statistically indistinguishable from zero (McAuliffe, 2017; Winking & Mizer, 2013).

We know of only two experiments that have addressed how the desire *to appear altruistically punitive* could masquerade as the desire *to be altruistically punitive* in the third-party punishment game. First, Pedersen, Kurzban, & McCullough (2013) adjusted the standard game so that (a) Adjusters could either punish or reward Dictators; and (b) Adjusters' behavioral options were not framed explicitly as opportunities to respond to Dictators' mistreatment of the Receiver. Instead, Adjusters' ability to take action against the Dictator was presented as a second

and entirely independent modified dictator game in which the Dictator could either give some of his or her own money to another player or else pay a cost to destroy some of that player's money (with no benefit to the Dictator from doing so). In this second dictator game, subjects were ostensibly randomly assigned to take the role of Dictator and the Dictator from Game 1 was ostensibly randomly assigned to take the role of Receiver. By temporally and logically separating Adjusters' discovery of the Dictator's stingy behavior in Game 1 from Adjusters' opportunity in Game 2 to punish the Dictator from Game 1, Pedersen et al. (2013) avoided establishing common knowledge that removing money was an opportunity to express disapprobation of the Receiver's unfair decision as Dictator in Game 1. Pedersen et al. also ran analogous conditions in which Receivers from Game 1 served as Dictators in Game 2 to evaluate whether unfairly treated Receivers spontaneously discovered and took advantage of their ability to punish the Dictator who had taken money from them in Game 1. Pedersen et al. found that the victims of the Dictator's behavior during Game 1 became angry toward the Game 1 Dictator and then punished him or her in Game 2 (i.e., there was second-party punishment). However, mere witnesses of the Game 1 Dictator's norm-violating behavior toward another player did not become angry or punish the Game 1 Dictator when they had the opportunity to do so in Game 2 (i.e., there was no third-party punishment.

A potential concern with Pedersen et al.'s (2013) approach to eliminating the common knowledge that monetary deductions represent punishment is that third parties might not have believed that the Receiver understood that the removal of money in Game 2 was intended as punishment for the Receiver's stinginess as Dictator in Game 1. If third parties condition their decisions to punish on whether doing so would deter future selfishness, the situation created by Pedersen et al. (2013) may have artificially reduced third-party punishment. This concern is

substantially mitigated by the fact that the actual victims of stingy Dictators did frequently punish them.

Additional reassurance comes from an experiment by Kriss, Weber, and Xiao (2016) that made the punishment opportunity common knowledge and still yielded qualitatively similar results to those of Pedersen et al. (2013). Kriss and colleagues modified Fehr and Fischbacher's (2004) third-party punishment game by telling subjects that their decisions to punish a norm-violator (at a cost to the punisher) would be executed only if subjects rolled an even number on a fair six-sided die. Subjects themselves reported the outcome of the die roll to the experimenter, which meant that subjects had the freedom to *communicate* a desire to punish norm-violators, but also the freedom to block the execution of that communicated desire by simply lying about the outcome of the die roll. Consistent with the idea that Adjusters commit to third-party punishment to signal their disapproval of norm violations rather than out of a desire to sanction the norm violation, witnesses who committed to paying money to punish a Dictator who had mistreated a third subject reported a much smaller percentage of even-numbered die rolls (21.7%) than would be expected by chance (50%). By contrast, the actual victims of the Dictator's stingy behavior reported a significantly *larger* percentage of even-numbered die rolls than what would be expected by chance, implying that some victims dishonestly *prevented their desire to punish from being thwarted* by an odd-numbered die roll. Moreover, the magnitude of punishment was not associated with self-reports of even die rolls, suggesting that harsher punishment does not correspond to a stronger desire to punishment. This finding is problematic for the view that the third-party punishment game measures punitive sentiment, for a measure is only valid if difference in measurement scores are caused by differences in the construct of interest (Borsboom, Mellenbergh, & van Heerden, 2004). Taken alongside Pedersen et al.'s (2013)

results, Kriss et al.'s (2016) findings suggest that the motivation underlying third-party punishment in the standard third-party punishment game is not a motivation *to be* altruistically punitive. Instead, it is a motivation *to appear* altruistically punitive.

Beyond "altruistic" punishment: Third-party punishment for personal benefit

Though these two recent experiments cast some doubt on the existence of a robust human propensity for altruistic third-party punishment, witnesses of harms obviously do sometimes take actions to punish harm-doers in real life (e.g., Phillips & Cooney, 2005). We propose that a primary function of third-party punishment—that is, a reason why a propensity for third-party punishment evolved during human evolution—is not to altruistically create benefits for strangers, but instead to *deter aggressors from harming individuals with whom the punisher shares a fitness interest* (Pedersen, McAuliffe, & McCullough, under review). On this view, the costs of third-party punishment can be offset via fitness benefits that accrue by deterring future harms toward victims in whom the punisher has a welfare stake (e.g., kin, mates, friends, and coalition members). People could plausibly estimate the interdependence of their welfare and others' welfare from a variety of fitness-relevant inputs (Roberts, 2005; Tooby & Cosmides, 2008; Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008), including ancestrally valid cues of genetic relatedness (e.g., sibship; Lieberman, Tooby, & Cosmides, 2007), past experiences of cooperative or exploitive interaction (Krasnow et al., 2012), horizons for future profitable interaction (Delton, Krasnow, Cosmides, & Tooby, 2011; McCullough, Pedersen, Tabak, & Carter, 2014), and shared parental investments (Clutton-Brock, 1989).

Three lines of evidence suggest that humans do, in fact, punish as third parties on behalf of victims with whom their welfare is interdependent. First, ethnographic evidence reveals that third parties punish those who have imposed costs upon their genetic relatives (Boehm, 1987;

Chagnon & Bugos, 1979; Ericksen & Horton, 1992). Second, Phillips and Cooney (2005) collected data on 136 recalled conflicts involving 852 third-party witnesses by interviewing men imprisoned for assault or homicide. Of the third parties with distant ties to one of the disputants (i.e., not a friend, family member, or fellow gang member), only 1% intervened in the conflict. In contrast, roughly 54% of third parties with individual ties (i.e., friends) and roughly 72% of third parties with group ties to one of the disputants (i.e., a family member or member of the same gang) did intervene (for related findings with children, see Huitsing & Veenstra, 2012).

Third, a closer examination of the elicitors of moral outrage, which has been proposed as part of the proximate motivational system that impels altruistic punishment (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002; Pedersen et al., 2013; Pedersen et al., under review; Petersen, Sell, Tooby, & Cosmides, 2010) can shed light on the conditions under which we can expect third parties to punish. Social psychologists have generally failed to find evidence that moral outrage exists in the absence of self-relevant concerns (reviewed in Batson, 2015). Instead, empathy for victims may be a requisite condition to experiencing anger on their behalf. For example, Batson, Kennedy, et al. (2007) found that third parties experienced anger upon witnessing another person mistreat a stranger only if they had been experimentally induced to feel empathy for the stranger prior to the mistreatment. Empathy is normally only felt for valued victims (Batson, Eklund, Chermok, Hoyt, & Ortiz, 2007), so Batson, Kennedy, et al.'s (2007) results imply that third-party anger is also typically felt only for valued victims. Such findings accord well with the proposal that empathy and anger are outputs of mechanisms that generate estimates of welfare interdependence (Pedersen et al., under review): On this view, people who perceive their welfare to be interdependent with a victim's experience empathy for the victim, anger toward the person who harmed the victim, and retaliatory motivation on the victim's behalf. Of course, we do not

mean to suggest that welfare interdependence is the only possible route to fitness benefits for third-party punishment. Indeed, other researchers have proposed that third-party punishment could produce reputational benefits for the punisher (Barclay, 2006, 2013; Kurzban et al., 2007; see also Halevy & Halali, 2015) and that it could deter future aggression toward the punisher (Krasnow et al., 2016; see also Petersen, Sell, Tooby, & Cosmides, 2010). Thus, our proposal here is merely that perceptions of welfare interdependence could play a key (though hardly singular) role in the psychological systems that regulate third-party punishment.

Third-party punishment in the context of insults and retaliatory aggression

If humans' propensities for third-party punishment evolved, in part, in response to selection pressure for obtaining fitness benefits by intervening to deter harms against those with whom the punisher has a shared fitness interest, we should find that (a) third-party punishment on behalf of strangers is rare in the absence of other potential benefits (e.g., the reputational benefits associated with being seen as a good steward of social norms), and (b) third-party punishment is sensitive to cues of welfare interdependence with a victim. Additionally, we should find that (a) third parties' anger and empathy on behalf of strangers is rare, and (b) anger and empathy are sensitive to cues of welfare interdependence with victims. To test these predictions, we designed five experiments in which we attempted to minimize the experimental demand that is typical of most third-party punishment experiments. To do so, we borrowed laboratory tasks from the social psychology literature on aggression, which provide unobtrusive and externally valid tools for detecting an anger-based motivation to impose retaliatory harm (Anderson & Bushman, 1997; Giancola & Chermack, 1998). In the experimental situation we devised, subjects were insulted by another subject in a way that has elicited anger and retaliatory aggression from victims in previous experiments (Bushman & Baumeister, 1998; Harmon-Jones

& Sigelman, 2001). We modified this task so that it involved three subjects—an insulter, the victim of the insult, and a third subject—so we could evaluate whether witnesses of insults impose punishment on insulters in the same manner as the victims of those insults typically do.

After either receiving an insult or witnessing the insult of the third subject, subjects completed an ostensibly unrelated task that gave them an opportunity to aggress against other subjects (insulters, victims, and other third-party witnesses) by setting the duration and volume of an irritating sound blast to which other subjects would be subjected (similar to Bushman & Baumeister, 1998). Intentional inflictions of physically aversive experiences are used frequently to measure punishment in social psychology (e.g., Bastian, Jetten, & Fasoli, 2011; Nelissen, 2012; Nelissen & Zeelenberg, 2009; Twenge, Baumeister, Tice, & Stucke, 2001). Although this approach to measuring punishment obviously differs from the standard experimental-economic approach, it bears the necessary and sufficient features of a prima facie valid measure of third-party punishment: the disinterested witness of a harm imposes retaliatory harm on the harm-doer.

Even so, third-party punishment has been studied almost exclusively with experimental-economic methods. So why would we use a measure of retaliatory aggression to study it here? First, because modern validity theory not only licenses it, but in fact demands it (Markus & Borsboom, 2013). Constructs exist independently of the methods used to measure them (Green, 1992), so if a construct is ontologically real, its existence must in principle be confirmable with multiple instruments whose methods are distinct (Campbell & Fiske, 1959). If a phenomenon identified with a single measure cannot be confirmed on a methodologically diverse set of instruments, researchers must remain open to the possibility that they have misidentified the construct or constructs that cause the scores on the original instrument. The fact that modifications to the standard third-party punishment game make third-party punishment all but

disappear (Kriss et al., 2016; Pedersen et al., 2013) attests to the fact that our validity concern is a live one. Without proper respect for the distinction between constructs and the methods used to identify those constructs, research on altruistic punishment will remain stuck in a limbo of uncertainty as to whether humans even possess such a motivation: What appears to be evidence of altruistic third-party punishment may instead be evidence of the human desire to *appear* altruistically punitive in others' eyes (Kriss et al., 2016; Pedersen et al., 2013). For this reason, research on altruistic punishment requires additional experiments that examine the concept using instruments that do not involve the methodology of experimental economics games.

The second reason we used a social situation that involved insults and retaliatory aggression is that it enabled us to study third-party punishment in response to violations of respect. Content validity is crucial for affirming the meaning and assessing the generalizability of results (Sireci, 1998); the fact that research on third-party punishment has focused almost exclusively on fairness violations limits scientists' ability to specify what the third-party punishment literature teaches us. To what real-life behaviors do results from the third-party punishment game enable us to make generalizations—punitive responses to inequity, or to violations of any social taboo? And if third-party punishment of fairness is an artifact—as, we contend, the results from Pedersen et al. (2013) and Kriss et al. (2016) suggest—do we then conclude that third-party punishment on behalf of strangers does not exist at all, or just that it does not occur in response to fairness violations? More generally, our focus on retaliatory aggression in response to insults enables us to build a theoretical bridge between the third-party punishment literature and a large cross-cultural literature on the social consequences of violated respect (Daly & Wilson, 1988; Haidt, Koller, & Dias, 1993; Nisbett & Cohen, 1996).

Finally, we note that our operationalization of punishment as a noise blast did not require subjects to pay costs in order to enact their desires to punish. This is a departure from the more customary procedures in this literature, which generally require punishers to pay money in order to enact their punishment decisions. We did not require subjects to pay costs precisely to encourage people to punish in accordance with their desires: In economic games, the less expensive punishment is, the more people punish (for reviews, see Guala, 2012; McCullough, Kurzban, & Tabak, 2013). Moreover, a comprehensive review of the ethnographic literature on punishment (Guala, 2012) revealed that punishment reliably occurs only when it is low-cost (in terms of time, physical threat, probability of reprisal, etc.). Therefore, the cost-free nature of punishment in our experiments should have maximized subjects' willingness to punish. Using a cost-free method of punishment, then, provides a generous test for the altruistic punishment hypothesis: if disinterested third parties do in fact possess a motivation to punish strangers who have harmed other strangers, even if they must pay costs to do so, then they should be even more willing to punish in such situations when they can do so free of charge.

The five experiments

We conducted five experiments to investigate these questions. The first of the five experiments was a hypothetical vignette study designed to shed light on the maximum extent to which subjects' propensities to engage in any altruistic third-party punishment could be motivated solely by a desire to communicate one's respect for moral norms and willingness to defend them. We asked subjects to imagine that they were in an experiment with two strangers and either (a) were insulted by one of the strangers or (b) witnessed one of the strangers insult the other one. Then, subjects reported how they thought they would feel and how they would

respond with the sound blast apparatus that we used for real-time responses in Experiments 2-5 (in which subjects were led to believe they were actually interacting with real people).

In Experiment 2, we tried to elicit real-time third-party punishment on behalf of a stranger via an empathy manipulation. In Experiment 3, we tried to elicit third-party punishment on behalf of a stranger via manipulations of perceptions of welfare interdependence. In Experiment 4, we tested whether third parties would punish on behalf of their friends by having subjects bring a friend with them to the experiment and then having those friends interact with two strangers. Then, subjects either (a) received an insult from a stranger, (b) witnessed a stranger insult the subject's friend, or (c) witnessed a stranger insult another stranger. In Experiment 5, we used four-person design similar to the set-up of Experiment 4 to test whether third parties would punish on behalf of strangers with an additional stranger present. Finally, we conclude this paper with a meta-analytic summary of the results of these five experiments. Throughout, we compare third-party punishment (punishment of the insulter by a witness of the insult) with second-party punishment (punishment of the insulter by a victim of the insult).

Experiment 1

To what extent could previous behavioral evidence for altruistic third-party punishment reflect a desire to depict oneself as concerned about observing and defending social norms? The behavioral propensities that people express in hypothetical vignettes (Aguinis & Bradley, 2014) are ideal for setting a theoretical upper bound on the strength of this desire because subjects' responses to hypothetical vignettes are the quintessence of cheap talk: costless to transmit, non-binding, and unverifiable, enabling subjects to portray themselves in any light they please without facing any consequences for doing so (Crawford & Sobel, 1982). In general, people's responses to vignettes conspicuously inflate their apparent tendencies to be outraged by

mistreatment of others. For instance, people respond to vignettes by indicating that they would be bothered by, and then sanction, racist comments when in real life they actually seem unfazed by them (Karmali, Kawakami, & Page-Gould, 2017; Kawakami, Dunn, Karmali, & Dovidio, 2009). More relevant to our concerns here, people respond to hypothetical vignettes that simulate the third-party punishment set-up of Pedersen et al. (2013) by indicating that they would be angered by, and then punish, a selfish dictator, even when laboratory results show that subjects who experience the situation in real time seem largely indifferent to it (Pedersen et al., 2013; see also Balafoutas & Nikiforakis, 2012).

To assess this possibility, we presented subjects with a hypothetical vignette that simulated the social situation that subjects actually experienced in real time in Experiments 2-5. We predicted that those subjects who imagined they received an insult and those subjects who imagined they witnessed a stranger receive an insult would both claim that they would become angry at, and punish, the insulter. Additionally, we predicted that subjects who imagined they witnessed a stranger receive an insult would claim that they would feel more empathy toward the victim than they would toward the insulter.

Method

Subjects

Subjects (N = 456; 198 female) were recruited from Amazon Mechanical Turk in exchange for $0.50 and were told the study involved "consider[ing] a situation in which you are participating in a psychology experiment and how you would respond and feel." Subjects participated online via SoPHIE (Software Platform for Human Interaction Experiments; Hendricks, 2012). All methods described in this paper were approved by the University of Miami's institutional review board. Data for all experiments are available at https://osf.io/62x9t/.

Procedure

Subjects read a hypothetical version of the social situation that we used in Experiments 2-5 (see supplemental material Appendix A for full transcript) and were randomly assigned to one of two conditions in which they imagined either: (a) receiving an insult or (b) witnessing a stranger receive the insult (Figure 1). They were told to imagine that they were completing a task that involved writing an essay about a personally important social issue, and then exchanging and reviewing essays over a computer network with two other individuals, Person A and Person B. All of the feedback on each person's essay was apparently reasonable and mildly positive, except for one review that Person A made either about Person B's essay or the participant's essay: "I can't believe that an educated person would think like this. I sincerely hope that this person learns a thing or two."

Self-reported emotional reactions. Subjects were then asked to report how they believed they would feel toward the two other interactants using the same emotion measures as Experiments 2-5. Instructions read "Please indicate the extent to which you would feel the following emotional response toward [person]." Subjects responded on a Likert-type scale from 0 (not at all) to 5 (extremely). Mixed among several distractor items were items that we used (as in Pedersen et al., 2013) to measure *anger* (mean of "angry," "mad," and "outraged") and *empathy* (mean of "compassionate," "empathic," and "sympathetic"). See Table 1 for reliabilities for all measures for all five experiments.

Punishment. Next, participants were asked to imagine that the experimenter wanted participants to test out a sound recording to gauge how participants in a future experiment would be likely to react. In the imagined scenario, the experimenter assigns the participant to decide how loud the sound will be and how long it will play for Person A and Person B, who will each

report how the sound made them feel. The sound is described is an irritating white noise that is similar to radio interference. Before continuing, participants heard a six-second sample of the sound. After reading and imagining this scenario, in randomized order, participants indicated on a scale from 1 (extremely quiet) to 10 (extremely loud) how loud they would make each person's sound sample. In the same order, participants also typed how many seconds they would have each person listen to the sound. A composite measure of punishment was created by taking the mean of the standardized values of both volume and duration (which was natural log-transformed due to skewness)[2] of the sound blast. These two values were moderately correlated for both Person A and Person B, *r*s = .64 and .57, *ps* < .001. The decision to create a composite variable was made *a priori* and was based on a similar decision by Bushman and Baumeister (1998), who reported a correlation of $r = .32$ between blast volume and duration.

Experiment 1 Results

Means and standard deviations for all major variables appear in Table 2. Throughout the paper, we conducted main analyses using Bayesian linear mixed models with Markov chain Monte Carlo (MCMC) estimation via the MCMCglmm package (Hadfield, 2010) in R version 3.1.2. Linear mixed models are more general than ANOVA approaches, enabling us to simultaneously model within-subject and between-subject effects (Barr, Levy, Scheepers, & Tily, 2013). We specified random intercepts for subjects in all models with within-subjects effects to account for non-independent observations. In such models, a Bayesian approach with MCMC estimation estimates *p*-values more accurately than do traditional approaches (Baayen, Davidson, & Bates, 2008; Bolker et al., 2009; Hadfield, 2010). We used non-informative priors in all Bayesian models and, unless otherwise noted, predictors in all models were dummy coded.

[2] We performed this transformation prior to testing our hypotheses and did so for every experiment in this paper.

In the text, we report Cohen's *d* effect sizes[3] and *p*MCMC (a Bayesian "*p*-value" based on MCMC estimation, defined as two times the probability that the parameter value is less than or greater than zero, using the smaller of these probabilities; Hadfield, 2010). In tables, we report parameter estimates (posterior means) and their associated 95% highest density intervals (HDIs; i.e., the 95% most credible values). Parameter estimates were considered statistically significant if both (a) their 95% HDI did not contain zero and (b) *p*MCMC < .05.

Did subjects' self-reported intention to punish vary as a function of whether the hypothetical victim was the subject or a stranger?

First, we evaluated the between-subjects comparison of the amounts of punishment that subjects claimed they would administer to insulters as a function of whether they imagined either that they received the insult or witnessed a stranger receive the insult (see Figure 2 and Table 3). Hypothetical victims of insults (M = .34, SD = 1.05) did not claim that they would punish more in absolute terms than did witnesses (M = .25, SD = .81), *p*MCMC = .266, d = .10.

Next, we focus on the within-subjects comparison of stated intention to punish the insulter *relative* to stated intention to punish the non-insulter (i.e., a within-condition control). Hypothetical victims of insults claimed they would punish insulters more than they would punish non-insulters (M = -.36, SD = .81), *p*MCMC < .001, d_z = .57. Likewise, hypothetical witnesses of insults claimed they would punish insulters more than they would punish non-insulters (M = -.24, SD = .72), *p*MCMC < .001, d_z = .54. To test whether self-reported intention to punish insulters, relative to non-insulters, varied between hypothetical victims and witnesses, we created punishment difference scores for each subject by subtracting self-reported intention to punish the

[3] Within-subjects *d*s, referred to as d_z, were calculated by dividing the mean of subjects' difference scores by the standard deviation of their difference scores (Lakens, 2013). Between-subjects *d*s, referred to as *d*, were calculated by dividing the mean difference of the independent groups by their pooled standard deviation.

non-insulter from self-reported intention to punish the insulter. Next, we ran a Bayesian linear regression predicting these difference scores with a dummy-coded variable for condition (victim vs. witness). The difference scores for hypothetical victims were significantly greater than those for hypothetical witnesses of insulted strangers (b = .21, pMCMC = .037). Thus, victims of insults forecasted that they would punish insulters, relative to non-insulters, to a greater extent than did witnesses of insults. Despite the statistical significance of this "difference in differences," their respective effect sizes were almost identical ($d_{z.}$ = .57 vs. d_z = .54).

Did forecasted anger vary as a function of whether the hypothetical victim was the subject or a stranger?

Hypothetical victims of insults (M = 1.98, SD = 1.36) stated that they would feel angrier toward insulters than did hypothetical witnesses of insults (M = 1.45, 1.31), pMCMC < .001, d = .41. Next, we evaluated the within-subjects comparisons of self-reported anger toward the insulter relative to anger toward the non-insulter. Hypothetical victims claimed that they would feel angrier toward insulters than they would toward non-insulters (M = .64, SD = 1.03), pMCMC < .001, d_z = .78. Likewise, hypothetical witnesses of insults claimed that they would feel angrier toward insulters than they would toward non-insulters (M = .48, SD = .84), pMCMC < .001, d_z = .67. To test whether self-reported anger toward insulters, relative to non-insulters, varied between hypothetical victims and hypothetical witnesses, we created anger difference scores and ran a regression in the same way we did for the punishment data. The difference scores for hypothetical victims were significantly greater than those for hypothetical witnesses of insulted strangers (b = .36, pMCMC = .016), indicating that hypothetical victims forecasted more anger toward insulters, relative to non-insulters, than did hypothetical witnesses.

Did hypothetical witnesses of insults forecast that they would feel empathy for victims?

Hypothetical witnesses of insults claimed they would feel more empathy for the victim (M = 2.27, SD = 1.21) than for the insulter (M = .83, SD = .91), pMCMC < .001, d_z = .88.

Experiment 1 Discussion

As predicted, subjects who imagined receiving an insult and subjects who imagined witnessing a stranger insult another stranger both claimed that they would (a) punish insulters more than they would punish non-insulters and (b) be angrier toward insulters than toward non-insulters. These results mirror results from third-party punishment games inasmuch as they suggest a human tendency to retaliate against norm-violators. However, analogously to what others have found (Karmali, Kawakami, & Page-Gould, 2017; Kawakami, Dunn, Karmali, & Dovidio, 2009; Pedersen et al., 2013), these results might also reflect cheap talk that does not accurately reflect how people actually respond to real-time experiences with unjustified insults. Thus, in Experiment 2, we created a laboratory situation in which subjects either directly suffered—or else merely witnessed—what they believed was a personal insult that had just occurred in real time. Additionally, we tested whether experimentally induced empathy for victims would lead witnesses to become angry at, and punish, insulters.

Experiment 2

Do third parties punish on behalf of strangers and does empathy increase anger and third-party punishment?

Experiment 2 was a real-time version of the social situation that Experiment 1 subjects considered hypothetically. Our goal was to test whether subjects would become angry at, and punish, insulters when subjects either (a) received an insult or (b) witnessed a stranger insult another stranger. We also added a second (though, unfortunately, unsuccessful) manipulation of

empathy in hopes of testing whether induced empathy for victims increases third-party punishment. We predicted that (a) victims of insults would become angry at, and punish, insulters, (b) witnesses of insults who did not receive an empathy induction would not become angry at, or punish, insulters, and (c) witnesses of insults who did receive an empathy induction would become angry at, and punish, insulters (Batson, Kennedy, et al., 2007).

Method

Subjects

Subjects (N = 147; 81 female) were undergraduates at the University of Miami who participated for partial course credit and $10. Before data analysis, we flagged all subjects who during debriefing expressed suspicion that their interactions with other subjects had been staged (n = 30). These exclusions left our final sample at N = 117 (67 female). To ensure that differential suspicion across experimental conditions did not influence our conclusions, we also ran intent-to-treat analyses on the full samples for Experiments 2-5 and report those results in the supplemental material. All qualitative differences between the results presented here and those from the intent-to-treat analyses (i.e., differences in significant effects) are denoted in the main text with footnotes. None of the results from the intent-to-treat analyses for any of the experiments yield substantively different conclusions from those of the analyses reported here.

Procedure

Subjects were seated at individual computers in private carrels and told they would be interacting with two other subjects—located either in the same room or in different rooms—over a computer network during their experimental sessions. To increase the believability of the interactions and promote subjects' memory of which subjects took each action, subjects were identified to each other by name. If fewer than three subjects showed up for a session, the session

was still run under the ruse that the other subjects were in other rooms; interactions with those subjects, who were referred to by randomly selected names, were fully staged. Subjects were randomly assigned to one of four conditions in a 2 (target of insult: self, stranger) by 2 (empathy manipulation: no empathy, empathy) design (see Figure 1).

Set-up for insult. We used a multi-person extension of the "insulting essay evaluation" that we used in hypothetical form in Experiment 1. When used in the laboratory, this paradigm reliably elicits anger (Bushman & Baumeister, 1998; Harmon-Jones & Sigelman, 2001). We slightly modified the topic for the essay and the corresponding insult to accommodate the empathy manipulation (see below). Subjects were given five minutes to type a short essay about something interesting that had recently happened to them (suggestions for possible topics were a recent accomplishment, a setback, or a strange experience) and were instructed that the quality of their writing would be judged by the other subjects in their session. They were told that all three subjects' essays would presently be circulated to all three subjects in their interaction group, who would each read and evaluate the essays. Once all three subjects finished their essays, subjects read the other two essays (presented in random order). After each essay, they provided a one- or two-sentence written review and rated the essay on several Likert-type scales (e.g., how unintelligent/intelligent the essay was on a scale from 1 to 9; these ratings were not analyzed, they were collected to add credibility to the insulting review [see below]). The essays we circulated were in fact written by the other subjects in the room; for sessions in which fewer than three subjects were present in the room, canned essays (written by undergraduates under the same conditions) were substituted. All interactions following the essays were fully staged.

Empathy manipulation. For the half of the subjects that were in the empathy condition, one of the essays that they read described a recent breakup that the writer had just experienced

and was suffering from. This essay has been used in several previous experiments to elicit empathy (e.g., Batson, Kennedy, et al., 2007). The other essay that subjects read, which was determined randomly, described either a birthday party or a geo-caching excursion. For the other half of subjects in the no-empathy conditions, these latter two essays were the two they read. For subjects in the empathy condition who were assigned to be insulted personally, the insulter's essay was always neutral and the non-insulter's essay was the empathy essay. For subjects in the empathy condition who were assigned to witness a stranger receive an insult, the insulter's essay was always neutral and the victim's essay was the empathy essay.

Insult. Next, subjects read the two other subjects' (bogus) evaluations of each of the essays (i.e., four reviews in total). The evaluations from the other subjects were slightly positive (e.g., "I like this essay and think the author wrote it pretty well."), except for one evaluation of one essay, which was insulting: "I can't believe an educated person would write like this. I hope this person learns something while at UM [University of Miami]." The insulted essay received lower ratings (M = 2.5) on the Likert-type items (e.g., unintelligent/intelligent) than did the rest of the essays (Ms = 7.5 to 8.2; see Figure 3 for a screenshot of the negative review).

After reading each evaluation, subjects rated how fair and how accurate they thought the evaluation was on scales from 0 (not at all) to 9 (totally). A composite fairness/accuracy score was created for both reviews of the essay that was insulted by taking the mean of these two items; composite scores were not created or analyzed for the reviews of the other essays). These ratings insured that subjects attended to the insulting evaluation and served as a manipulation check to confirm that the insult was perceived as unfair (relative to the other reviews).

Self-reported emotional reactions. Subjects then completed the same self-report emotion measures as in Experiment 1, though with the instructions edited to read "Please indicate the extent to which you are feeling the following emotional response toward [person]."

Sound blast ("punishment"). Next, subjects were told that the main portion of the experiment was finished but that the experimenters were evaluating various sounds for use in future experiments and needed some feedback on how pleasant or unpleasant the sounds were. In reality, this was a cover story that provided subjects the opportunity to administer punitive sound blasts to the other subjects while minimizing experimental demand for punishment (Weber & Cook, 1972). Subjects were then told they had been randomly chosen to be an "audio administrator" and would therefore be asked to assign sound samples to the other subjects, who had been assigned to be "sound raters." Subjects listened through headphones to three short samples, of different volumes, of an unpleasant static sound and were asked to rate how pleasant/unpleasant it was. Next, they assigned a volume level from 1 (quietest) to 10 (loudest), and then held down the space bar for as long as they wished (to assign a duration) for each of the other subjects individually. Subjects were led to believe the sound was playing in real time for the other subjects. In reality, no actual sound stimulus was delivered to other subjects. A composite measure of punishment was created in the same way as in Experiment 1.

Experiment 2 Results

Results

Means and standard deviations for all major variables appear in Table 4. Analyses were conducted in the same manner as for Experiment 1. See tables 5-7 for full model results.

Manipulation Check: Did empathy for victims vary as a function of the empathy manipulation?

First, we evaluated whether witnesses of insults experienced empathy for victims as a function of the empathy manipulation. Witnesses in the empathy condition did not report more empathy toward victims (M = 2.59, SD = 1.68) than did witnesses in the no-empathy condition (M = 2.46, SD = 1.22), b = -.13, pMCMC = .714, d = .09. Furthermore, witnesses in the empathy condition did not report more empathy for victims than victims of insults reported for the non-insulter in the empathy condition (M = 2.61, SD = 1.57), b = .02, pMCMC = .949, d = .01, indicating that the empathy manipulation did not increase empathy toward victims. Thus, our manipulation of empathy failed, undermining our ability to test our hypothesis of the causal role of empathy in punishment and anger. Nevertheless, we were still able to test for differences in punishment and anger between victims and witnesses. To maximize our power in doing so, all subsequent analyses are collapsed across empathy conditions.

Manipulation check: Did subjects perceive the insulting review as unfair/inaccurate?

Subjects who received an insulting review of their own essay reported that the review was less fair/accurate (M = 2.39, SD = 2.70) than was the non-insulting (i.e. slightly positive) review of their essay (M = 7.39, SD = 1.64), pMCMC = < .001, d_z = 1.36. Likewise, witnesses of insults reported that the insulting review was less fair/accurate (M = 3.44, SD = 2.69) than was the non-insulting review that the victim wrote for the insulter's essay (M = 7.23, SD = 7.23), pMCMC < .001, d_z = 1.27. Interestingly, there was a small but statistically significant review by target-of-insult interaction (see Table 6): Subjects who were personally insulted rated the insulting review as less fair/accurate than did subjects who read an insulting evaluation of

another subject's essay, *p*MCMC < .013, *d* = .39, whereas recipients and witnesses of insults did not differ in their ratings of the non-insulting review, *p*MCMC = .695, *d* = .09.

Did punishment vary as a function of whether the victim was the subject or a stranger?

Next, we evaluated the between-subjects comparison of the amount of punishment that subjects administered to insulters as a function of whether they received the insult or witnessed a stranger receive the insult (see Figure 4). Victims punished insulters (M = .35, SD = .96) more in absolute terms than did witnesses of insults (M = -.01, SD = .86), *p*MCMC = .030, *d* = .39[4].

Next, we evaluated the within-subjects comparison of punishment of the insulter *relative* to punishment of the other person. There was a within-subjects effect for victims of insults: Victims punished insulters more than they punished non-insulters (*M* = -.20, *SD* = .80), *p*MCMC < .001, d_z = .57. In contrast, this within-subjects effect was not significant for witnesses of insults: Witnesses of insults did not punish insulters more than they punished the victims of the insults (*M* = -.12, *SD* = .82), *p*MCMC = .364, d_z = .12. To test whether these differences in punishment of insulters, relative to non-insulters, varied between victims and witnesses, as in Experiment 1 we created punishment difference scores and ran a regression with a dummy-coded variable for condition (victim vs. witness).The difference scores for victims were significantly greater than those for witnesses of insulted strangers (*b* = .43, *p*MCMC =.016), indicating that victims punished insulters, relative to non-insulters, to a greater extent than did witnesses of insults. Thus, overall, victims of insults punished insulters whereas witnesses of insults did not, and the difference between victims and witnesses was statistically significant.

[4] When suspicious subjects were included in this analysis, the effect became marginally significant: *p*MCMC = .057, *d* = .30 (see Table S4).

Did anger vary as a function of whether the victim was the subject or a stranger?

First, we evaluated the between-subjects comparison of the amount of anger toward insulters that subjects reported as a function of whether they received the insult or witnessed a stranger receive the insult. Victims of insults reported more anger toward insulters (M = 2.16, SD = 1.59) than did witnesses of insults (M = .74, SD = .92), pMCMC < .001, d = 1.11.

Next, we evaluated the within-subjects comparison of anger toward the insulter *relative* to anger toward the other person. Victims of insults reported significantly more anger toward insulters than they did toward non-insulters (M = .54, SD = .92), pMCMC < .001, d_z = .90. Witnesses of insults also reported slightly more anger toward insulters than they did toward non-insulters, but the 95% HDI for this parameter estimate (just barely) included zero (M = .36, SD = .77), pMCMC = .055, d_z = .35. To test whether anger toward insulters, relative to non-insulters, varied between victims and witnesses, we again created anger difference scores and ran a regression. The difference scores for victims were significantly greater than those for witnesses of insulted strangers (b = 1.24, pMCMC < .001), indicating that victims of insults reported more anger toward insulters, relative to non-insulters, than did witnesses of insults. Thus, overall, victims of insults reported more anger toward the insulter relative to the non-insulter than did witnesses of insults, who reported marginally more anger toward insulters than toward non-insulters (i.e., toward the individuals who were the victims of the insult).

Did differences in anger mediate differences in punishment for victims and witnesses?

We were interested in whether victims punished insulters to a greater extent than did witnesses because victims became angrier at insults than witnesses did. To explore this possibility, we conducted a mediation analysis with condition (victims vs. witnesses) directly predicting punishment difference scores and indirectly predicting punishment difference scores

via anger difference scores. Thus, the analysis tested whether anger difference scores mediated the relationship between condition and punishment difference scores. This analysis was run using the PROCESS macro for SPSS (Hayes, 2013). A bias-corrected bootstrap confidence interval was used to evaluate the statistical significance of the indirect effect of the target of the insult on relative punishment through relative anger. In this mediation model, the indirect effect was significant b = .37, 95% CI = [.21, .60]. Thus, victims punished insulters more than non-insulters to a greater extent than witnesses did in part because victims were also angrier at insulters (relative to non-insulters) than witnesses were.

Did witnesses report empathy for insulted strangers?

Finally, witnesses of insults reported significantly more empathy toward victims (M = 2.12, SD = 1.58) than they did toward insulters (M = .84, SD = 1.09), pMCMC < .001, d_z = .68.

Experiment 2 Discussion

Our aims in Experiment 2 were (a) to test whether victims and witnesses would become angry at, and punish, insulters in a real-time experiment and (b) to test the causal role of empathy in third-party punishment by manipulating empathy toward victims of insults. Although our empathy manipulation had no noteworthy significant effects, undermining our ability to test the latter conjecture, Experiment 2 nevertheless yielded five important results. First, victims of insults punished insulters significantly more than they punished non-insulters. Second, witnesses of insults *did not* punish insulters any more than they punished non-insulters. Third, victims experienced more anger toward the insulter than they did toward non-insulters. Fourth, witnesses experienced slightly more anger toward the insulter than they did toward non-insulters, though the 95% HDI around the parameter estimate for this effect also included a value of zero. Fifth, a

mediation analysis suggested that victims punished insulters (relative to non-insulters) more than witnesses did specifically because they experienced more anger at insulters than witnesses did.

Overall, victims in Experiment 2 behaved as the subjects in Experiment 1 claimed they would react in similar circumstances: In response to receiving an actual insult, subjects were more punitive and angry toward insulters than toward non-insulters. In contrast, the results for witnesses in Experiment 2 diverge sharply with how subjects in Experiment 1 claimed they would react: In response to witnessing an actual insult in Experiment 2, subjects were no more punitive, and only marginally angrier, toward insulters than toward non-insulters; in Experiment 1, both of these effects were sizeable. However, witnesses in Experiment 2 did report more empathy for victims than they did for insulters, which is in line with how subjects in Experiment 1 claimed they would feel. Taken together, the results of Experiments 1 and 2 lend credence to our conjecture that people's beliefs in their own penchant for third-party punishment on behalf of strangers is due primarily to experimental demand and a desire to appear altruistically punitive.

Given the failure of the empathy manipulation, Experiment 2 did not equip us to discern empathy's causal role in third-party punishment. Consequently, we theorized that it may be possible to induce third-party anger, empathy, and punishment by manipulating their hypothesized common cause—perceptions of welfare interdependence (i.e., how much subjects perceive their own welfare as being contingent upon the welfare of another; Pedersen et al., under review; Roberts, 2005; Tooby & Cosmides, 2008; Tooby et al., 2008). Thus, in Experiment 3, we tried to manipulate estimates of welfare interdependence directly rather than by manipulating their hypothesized downstream effects on emotion systems. This manipulation, we hoped, would allow us to test whether estimates of welfare interdependence indeed regulate third-party punishment via their effects on anger and empathy.

Experiment 3

Do partner generosity and prospect of future interaction affect anger and third-party punishment?

In Experiment 3, we experimentally manipulated two cues that we expected to raise subjects' estimates of welfare interdependence with a partner who was initially a stranger. The first of these cues was the partner's generosity toward the subject during initial cooperative interactions. Everything else equal, the more benefits a partner provides, the more inherently valuable they are to one's own welfare which, in turn, should lead one to value the partner's welfare more than that of a less generous person (Roberts, 2005; Tooby et al., 2008). The second cue we manipulated was the prospect of future cooperative interaction with the partner following the initial cooperative interactions. Due to the prospect of the long-term benefits to be gained from cooperation over repeated interactions, cooperative partners are more valuable to one's welfare the more certain it is that productive interactions will continue (Axelrod & Hamilton, 1981; Delton et al., 2011; McCullough et al., 2014; Nelissen, 2014; Trivers, 1971). Following these manipulations, we used the insulting essay paradigm from Experiments 1 and 2 to create a situation in which the partner was harmed by another stranger, which was then followed by an opportunity for the subject to punish the harm-doer. Unlike in Experiments 1 and 2, subjects were not randomly assigned as either witnesses or victims of insults in this experiment; instead, subjects were always witnesses to insults. We predicted that both manipulations would increase punishment of insulters, anger toward insulters, and empathy toward victims. In line with results from Experiment 2, we also predicted that subjects would not become angry at, or punish, insulters when partner generosity and the probability of future interaction were low.

Experiment 3 Method

Subjects

Subjects (N = 250; 136 female) were undergraduate students at the University of Miami who participated for partial course credit and $10. Before data analysis, we flagged all subjects who during debriefing expressed suspicion that their interactions with other subjects had been staged (n = 44). These exclusions left our final sample at N = 206 (117 female).

Procedure

As in Experiment 2, subjects were run in small groups and were led to believe that they were interacting with two other subjects over computers. In reality, they interacted with sham computer partners. Subjects first played an iterated Trust Game (Berg, Dickhaut, & McCabe, 1995) with one of the sham partners after having been told that the amount of money they made in the game would be theirs to keep. In the Trust Game, an "Investor" starts with an endowment of money and is given the chance to transfer some of it to the "Trustee." Transferred money is quadrupled, and the Trustee can then return none, some, or all of the total back to the Investor. Subjects, as far as they were aware, were randomly assigned to the Investor role, and their partner was assigned to the Trustee role, for three rounds of play (the number of rounds to be played was not specified in advance to avoid end of game effects). Subjects were given $1.50 to use in each round of the Trust Game, and any amount the subject transferred to the partner in each round was be multiplied by 4. Thus, a $1 transfer became $4 in the partner's account. After the Investor's transfer, the partner could [ostensibly] back-transfer to the subject any amount (up to his total current holdings) he or she chose to.

We randomly assigned subjects to one of the six cells in a 3 (partner generosity: fair, generous, very generous) by 2 (prospect of future cooperative interaction: low, high) between-subjects design (see Figure 5).

Manipulation of perceived generosity. The partner's generosity was manipulated by having him return either 200% ("fair"; a 50/50 split), 250% ("generous"), or 300% ("very generous") of the subject's investment in each of the three rounds (i.e., the partner always returned the same percentage).

Manipulation of the prospect of future interaction. After the three rounds of the Trust Game, subjects were informed that they would either (a) resume playing the same iterated Trust Game with their partner toward the end of the session (high prospect of future interaction) or (b) not play any additional economic games with their partner (low prospect of future interaction).

Additional procedures. Following the manipulations, the rest of the procedure was nearly identical to Experiment 2, except that the essay topic and the insult were changed back from our slightly-altered versions in Experiment 2 to match those used previously in the literature (Bushman & Baumeister, 1998; Harmon-Jones & Sigelman, 2001). Specifically, subjects were given five minutes to type a short essay about any personally important issue they wished (suggestions for possible topics were abortion, gay marriage, marijuana legalization, healthcare, or alcohol laws). As in Experiment 2, the (sham) evaluations from the other subjects were slightly positive (e.g., "I can understand why a person would think like this."), except for one evaluation of one essay, which was insulting: "I can't believe an educated person would think like this. I hope this person learns something while at UM [University of Miami]." The subject's partner in the trust game was the recipient of the insult, whereas the insulter was a person with whom the subject had no prior interaction.

Experiment 3 Results

Means and standard deviations for all major variables appear in Table 8. Main analyses were conducted as Experiments 1 and 2, except that predictors were effect coded rather than dummy coded for ease of interpretation. Additionally, subjects were not victims of insults in this experiment, they were only witnesses. See Tables 9 and 10 for full model results.

Manipulation check: Perceived fairness/accuracy of the insulting review.

Subjects reported that the insulting review was less fair/accurate (M = 3.27, SD = 2.38) than the non-insulting review that the victim wrote for the insulter's essay (M = 7.50, SD = 1.61), pMCMC < .001, d_z = 2.39. The generosity and future interaction manipulations did not affect ratings of fairness/accuracy either individually or jointly, pMCMCs > .086 (see Table 9).

Did punishment vary as a function of the future interaction and partner generosity manipulations?

The generosity and future interaction manipulations did not affect punishment individually or jointly, pMCMCs > .105[5]. Furthermore, subjects did not punish insulters (M = .003, SD = .86) any more than they punished victims of insults (M = -.003, SD = .84), pMCMC = .97, d_z = .01 (see Figure 6).

Did anger vary as a function of the future interaction and partner generosity manipulations?

The generosity and future interaction manipulations also did not affect anger individually or jointly, pMCMCs > .070. However, subjects did report more anger toward insulters (M = .82, SD = 1.12) than they did toward victims of insults (M = .39, SD = .79), pMCMC < .001, d_z = .36.

[5] When suspicious subjects were included in this analysis, one term in the model became significant (pMCMC = .02), suggesting slightly less punishment of the insulter when the generosity level was a "fair" 50/50 split (see Table S7). However, given that the 95% HDI was very close to including zero [-.15, -.01] and the effect was not present in non-suspicious subjects, the effect does not appear robust and we did not investigate further.

Did empathy vary as a function of the future interaction and partner generosity manipulations?

Finally, the generosity and future interaction manipulations did not affect empathy individually or jointly, *p*MCMCs > .068. Subjects did report more empathy for victims of insults (M = 2.02, SD = 1.32) than for insulters (M = 1.47, SD = 1.21), *p*MCMC < .001, d_z = .41.

Experiment 3 Discussion

Experiment 3 yielded three main results. First, as in Experiment 2, we found no evidence of third-party punishment on behalf of strangers: witnesses of an insult did *not* direct a significantly greater amount of punishment toward the insulter than toward the victim of the insult. Second, witnesses *did* report significantly more (about one-quarter of a standard deviation) anger toward insulters than toward victims. Third, as in Experiment 2, witnesses reported significantly more empathy toward victims than they reported toward insulters.

Our main goal in Experiment 3, however, was to manipulate cues of welfare interdependence so that we could test their causal role in third-party punishment. In a sense, we aimed to create friendship in the lab. Our manipulations of partners' generosity and prospect of future interaction did not have any noteworthy effects on our dependent variables, however. Two explanations seem plausible. First, it might be the case that generosity and prospect of future interaction do not, in fact, affect estimates of welfare interdependence as theorized. Second, the manipulations simply might have been inadequate: Perhaps additional trials of the trust game or more potent manipulations would have had the effect we had intended. In Experiment 4, instead of trying to create friendship in the lab, we simply had subjects bring a friend with them to test whether third parties do indeed punish on behalf of their friends.

Experiment 4

Do third parties punish on behalf of their friends?

Although theory and observational studies suggest that friends play important roles in interpersonal conflicts (for review, see Frey, Pearson, & Cohen, 2015), we are aware of no laboratory experiment to date that has tested whether third-party witnesses to harms directed toward their friends punish their friends' harmdoers. To address this gap in the literature, we designed an experiment in which subjects either (a) were insulted by a stranger; (b) witnessed a stranger receive an insult from another stranger; or (c) witnessed their friend receive an insult from a stranger. We predicted that victims would become angrier and more punitive than would the witnesses of insulted friends and insulted strangers would. Because third parties' welfare is intrinsically more interdependent with their friends than with strangers, we also predicted that witnesses of insulted friends would punish insulters, report more anger toward insulters, and report more empathy toward victims than would witnesses of insulted strangers.

Experiment 4 Method

Subjects

Subjects (N = 222; 121 female) were undergraduates at the University of Miami who came with a friend to the lab. Both subjects in each friend pair participated for $10 and partial course credit (if enrolled in an introductory psychology course). Before data analysis, we flagged subjects who during debriefing expressed suspicion that their interactions with other subjects had been staged (n = 23). These exclusions left our final sample at N = 199 (112 female).

Procedure

Subjects were run in small groups and told they would be interacting over the computer network with three other subjects, one of whom was their friend. The other two subjects came to

the experiment by themselves and did not know any of the other subjects (they were recruited for Experiment 5, which was run simultaneously, did not contribute data to Experiment 4, and produced data that were statistically independent of the data analyzed for Experiment 4). To prevent subjects from inferring that the two subjects they did not know were also a pair of friends, a prompt was displayed at the beginning of the computer program that revealed that only the subject and the subject's friend knew each other. To increase the believability of the interactions and promote subjects' memory of which other subjects took what actions, subjects were identified to each other by name. If fewer than two stranger subjects showed up for a session, the session was still run under the guise that the stranger subject(s) were in other rooms and interactions with those subjects were fully staged; pairs of friends were always located in the same room and prevented from seeing or hearing each other during the session.

Subjects were randomly assigned to one of three conditions: They either (a) received an insult from a stranger; (b) witnessed a stranger receive an insult from another stranger; or (c) witnessed their friend receive an insult from a stranger (see Figure 7). Hence, when subjects themselves were insulted, there was an "insulter," a "non-insulter", and a "friend;" when subjects witnessed a stranger receive an insult, there was an "insulter," a "victim" and a "friend;" and when subjects witnessed a friend receive an insult, there was an "insulter", a "victim" (the friend), and a "non-insulter." The insulter was always a stranger. As in Experiments 2 and 3, subjects read two reviews for each essay, which lead to eight reviews total in this experiment. The rest of the procedure was identical to the post-manipulation procedures of Experiment 3.

Experiment 4 Results

Means and standard deviations for all major variables appear in Table 11. Main analyses were conducted as in Experiments 2 and 3. See Tables 12 and 13 for full model results.

Manipulation check: Did subjects perceive the insulting review as unfair/inaccurate?

As in Experiments 2 and 3, subjects in all conditions rated the insulting review as less fair/accurate than the non-insulting review. Subjects who received an insulting review of their own essay reported the review was less fair/accurate (M = 2.36, SD = 2.47) than was the non-insulting review of the essay (M = 7.72, SD = 1.20), pMCMC < .001, d_z = 1.92. Likewise, witnesses of insulted strangers reported as well that the insulting review was less fair/accurate (M = 3.54, SD = 2.57) than was the non-insulting review (M = 7.71, SD = .94), pMCMC <.001, d_z = 1.59. Witnesses of insulted friends also reported the insulting review was less fair/accurate (M = 2.19, SD = 2.0) than the non-insulting review (M = 7.56, SD = 1.19), pMCMC <.001, d_z = 2.15. As we found in Experiment 2, there was a significant review by victim-of-the-insult interaction (see Table 12): Subjects who were personally insulted rated the insulting review as less fair/accurate than did subjects who read an insulting evaluation of a stranger's essay, pMCMC < .001, d = .47, and subjects who read an insulting evaluation of their friend's essay also rated the insulting review as less fair/accurate than did subjects who read an insulting evaluation of a stranger's essay, pMCMC < .001, d = .58. Victims and witnesses of friends receiving insults did not differ in their ratings of the fairness/accuracy of the insulting essay, pMCMC = .614, d = .07. Ratings of the non-insulting review did not differ between victims and witnesses of insulted strangers (pMCMC = .971, d = .01), between victims and witnesses of insulted friends (pMCMC = .632, d = .13), or between witnesses of insulted friends and witnesses of insulted strangers (pMCMC = .655, d = .13). Thus, both victims and witnesses of insulted friends rated the insulting review as less fair/accurate than did witnesses of insulted strangers, and ratings of the fairness/accuracy of the non-insulting review did not differ among subjects.

Did punishment vary as a function of whether the victim was the subject, the subject's friend, or a stranger?

To test whether punishment varied as a function of whether the victim was the subject, the subject's friend, or a stranger, we evaluated the between-subject comparisons of the amounts of punishment that subjects administered to insulters as a function of whether they received the insult, witnessed a stranger receive the insult, or witnessed a friend receive the insult (see Figure 8). Victims punished insulters more *in absolute terms* than witnesses of insulted strangers did, *p*MCMC = .035, *d* = .36[6], but not more than witnesses of insulted friends did, *p*MCMC = .091, *d* = .29. Further, witnesses of insulted strangers and witnesses of insulted friends did not differ in their absolute punishment of the insulter, *p*MCMC = .663, *d* = .09. Thus, victims punished insulters more than witnesses of insulted strangers did, and witnesses of insulted friends punished insulters with an intensity between that of victims and witnesses of insulted strangers.

Next, we evaluated the within-subjects comparisons of punishment of the insulter *relative* to punishment of the other two people, which allowed for within-condition control comparisons. As Figure 8 shows, victims of insults punished insulters (*M* = .33, *SD* = .95) more than they punished non-insulters (*M* = -.28, *SD* = .76) or their friends (*M* = -.18, *SD* = .97), *p*MCMC < .001, d_z = .64 and *p*MCMC < .001, d_z = .44, respectively. Additionally, witnesses of insulted friends punished insulters (*M* = .08, *SD* = .74) more than they punished non-insulters (*M* = -.20, *SD* = .68), *p*MCMC = .012, d_z = .42, but not more than they punished their friends (*M* = .01, *SD* = 1.01), *p*MCMC = .527, d_z = .07. In contrast, witnesses of insulted strangers *did not* punish insulters (*M* = .02, *SD* =.76) more than victims (*M* = .07, *SD* = .68), *p*MCMC = .632, d_z = .08 or their friends (*M* = .18, *SD* = .92), *p*MCMC = .157, d_z = - .19. Thus, both victims and witnesses of

[6] When suspicious subjects were included in this analysis, *p*MCMC = .099, *d* = .26 (see Table S10).

insulted friends punished insulters more so than they punished non-insulters, whereas witnesses of insulted strangers did not.

To test whether these between-subjects differences in punishment of insulters relative to non-insulters (victims vs. witnesses of insulted friends vs. witnesses of insulted strangers) were reliable, we created punishment difference scores for each subject by subtracting punishment of the non-friend[7] from punishment of the insulter (as we did for Experiments 1 and 2). Next, we ran Bayesian linear regressions predicting the punishment difference scores with dummy-coded variables for condition (e.g., victim vs. witness of insulted friend). The difference scores for victims were significantly greater than those for witnesses of insulted friends (b = .32, pMCMC = .018) and witnesses of insulted strangers (b = .66, pMCMC < .001). That is, the amount that victims punished insulters compared to non-insulters was greater than the amount that witnesses of insulted friends or witnesses of insulted strangers punished insulters compared to non-insulters. Additionally, the difference scores for witnesses of insulted friends were significantly greater than those for witnesses of insulted strangers (b = .33, pMCMC = .014), indicating that the amount witnesses of insulted friends punished insulters compared to non-insulters was greater than the amount witnesses of insulted strangers punished insulters compared to non-insulters. According to the effect sizes associated with the difference in punishment of the insulter relative to punishment of the non-insulter, victims (d_z = .64) punished insulters about 8 times as harshly, and witnesses of insulted friends (d_z = .42) punished insulters about 5 times as harshly, as did witnesses of insulted strangers (d_z = .08). Thus, compared to non-insulters,

[7] We focused here on the difference score between insulters and non-friends because friends were "punished" at statistically identical levels to insulters in the two witness conditions, which possibly resulted from subjects using the sound blast apparatus on their friends without any retaliatory intent (perhaps to play a joke on their friends that they could discuss later). The data do not allow for a formal test of this explanation, but the lack of anger toward friends in those conditions (see below) suggests that punishment of friends was not driven by anger.

victims of insults punished insulters more than did witnesses of insulted friends or witnesses of insulted strangers, and witnesses of insulted friends punished insulters more than did witnesses of insulted strangers. Furthermore, witnesses of insulted strangers did not punish insulters significantly more so than they punished non-insulters.

Did anger vary as a function of whether the victim was the subject, the subject's friend, or stranger?

To address this question, we first evaluated the between-conditions (i.e., victim vs. witness of insulted friend vs. witness of insulted stranger) comparisons of the anger that subjects felt toward insulters as a function of whether they received the insult, witnessed a stranger receive the insult, or witnessed a friend receive the insult. Victims of insults reported more anger toward insulters *in absolute terms* than did witnesses of insulted strangers, pMCMC < .001, d = .97, and more than did witnesses of insulted friends, pMCMC < .001, d = .63. Witnesses of insulted friends also reported more anger toward the insulter than did witnesses of insulted strangers, pMCMC = .007, d = .38. Thus, victims reported more anger toward insulters than did witnesses of insulted friends who, in turn, reported more anger toward insulters than did witnesses of insulted strangers.

Next, we evaluated the within-subjects comparisons of anger toward the insulter *relative* to anger toward the other two people. As Figure 8 shows, victims of insults reported more anger toward insulters (M = 2.00, SD = 1.73) than toward non-insulters (M = .36, SD = .86), pMCMC < .001, d_z = .89, or toward friends (M = .15, SD = .67), pMCMC < .001, d_z = 1.07. Similarly, witnesses of insulted friends reported more anger toward the insulter (M = 1.04, SD = 1.26) than toward non-insulters (M = .29, SD = .58), pMCMC <.001, d_z = .60, or toward friends (M = .13, SD = .35), pMCMC < .001, d_z = .76. Finally, witnesses of insulted strangers reported more anger

toward insulters (M =.60, SD = 1.02) than toward victims (M = .20, SD = .53), pMCMC = .009, d_z = .36 or toward friends (M = .14, SD = .53), pMCMC = .003, d_z = .43.

Did differences in anger mediate differences in punishment for victims, witnesses of insulted friends, and witnesses of insulted strangers?

We were also interested in whether the between-condition (i.e., victim vs. witness of insulted friend vs. witness of insulted stranger) differences in anger toward insulters versus non-insulters mediated the differences in the between-condition differences in punishment of insulters versus non-insulters. To explore this possibility, we conducted three mediation analyses using the same method as in Experiment 2. For victims, the indirect effect was significant in the model comparing them to witnesses of insulted strangers (b = .34, SE = .10, 95% CI = [.18, .57], as well as in the model comparing them to witnesses of insulted friends, b = .21, SE = .08, 95% CI = [.09, .41]. The indirect effect for the model comparing witnesses of insulted friends to witnesses of insulted strangers was nearly statistically significant, though the effect size was relatively small, b = .07, SE = .05, 95% CI = [-.004, .19]. Thus, victims of insults punished insulters, relative to non-insulters, to a greater extent than did witnesses of insulted friends or witnesses of insulted strangers at least in part because they had greater anger toward insulters, relative to non-insulters, than did witnesses of insulted friends or insulted strangers. Additionally, witnesses of insulted friends punished insulters, relative to non-insulters, more than did witnesses of insulted strangers at least in part because of greater anger toward insulters, relative to non-insulters, although this mediation was relatively small in magnitude and the 95% confidence interval for the latter indirect effect (just barely) contained 0.

Did empathy vary as a function of the victim's identity?

To address this question, we first evaluated the between-subjects comparison of empathy toward victims as a function of whether they witnessed a friend or a stranger receive the insult. Witnesses of insulted friends reported more empathy for victims (i.e., their friends) than witnesses of insulted strangers reported for victims, *p*MCMC = .001, *d* = .56.

Next, we evaluated the within-subjects comparisons of empathy toward the victim relative to empathy toward the other two people. Witnesses of insulted friends reported more empathy toward friends (*M* = 2.23, *SD* = 1.31) than toward either non-insulters (*M* = 1.49, *SD* = 1.32), *p*MCMC < .001, d_z = .42, or insulters (*M* = 1.13, *SD* = 1.06), *p*MCMC < .001, d_z = .74. In contrast, witnesses of insulted strangers did not report more empathy toward victims (*M* = 1.49, *SD* = 1.32) than toward insulters (*M* = 1.24, *SD* = 1.01), *p*MCMC = .148, d_z = .19, and they reported less empathy for the victim than for friends (who were not insulted; *M* = 2.12, *SD* = 1.42), *p*MCMC < .001, d_z = -.45. However, the absolute magnitude of empathy for friends did not vary as a function of victim identity (pMCMCs > .64, *d*s < .08), suggesting that higher ratings of empathy toward friends when they were insulted did not result from the insult itself, but from higher empathy ratings toward friends more generally. In sum, witnesses of insulted strangers *did not* report more empathy for victims than for insulters, but witnesses of insulted friends did report more empathy for their friends than for insulters or non-insulters, probably because of a generalized tendency to empathize more with friends than with strangers.

Experiment 4 Discussion

Experiment 4 yielded six main results. First, victims of insults punished insulters more than did witnesses of insulted friends or of insulted strangers, and witnesses of insulted friends punished insulters more than did witnesses of insulted strangers. Second, as in Experiments 2

and 3, witnesses of insulted strangers did not punish insulters any more than they punished non-insulters. Third, victims of insults reported the most anger toward insulters, followed by witnesses of insulted friends, followed by witnesses of insulted strangers (who reported significantly more anger toward insulters than toward non-insulters). Fourth, we found evidence that differences in anger partially explained differences in punishment of insulters among victims, witnesses of insulted friends, and witnesses of insulted strangers. Fifth, subjects reported more empathy toward their friends than toward victims, insulters, or uninvolved bystanders, but empathy toward friends did not vary as a function of victim identity. Finally, in contrast to Experiments 2 and 3, witnesses of insulted strangers did not report more empathy for victims than they reported for insulters. Taken together, this pattern of results suggests that people experience an anger-based inclination to punish people who have insulted them, and a slightly weaker anger-based inclination to punish people who have insulted their friends, but no analogous inclination to punish people who have insulted strangers. These findings support our hypotheses that third-party punishment is in part driven by perceptions of welfare interdependence with the victim. Of course, the null results from Experiment 3 may speak against this interpretation, though it is impossible to tell whether it was simply a result of a failed manipulation of welfare interdependence. Further experiments are needed to distinguish these possibilities.

In Experiment 5, we extended the four-person paradigm from Experiment 4 to test whether third parties would punish on behalf of strangers not when their friend was present, but instead, when another stranger—also an uninvolved bystander—was present. The addition of an uninvolved bystander might tend to increase subjects' perception that they could receive approbation for engaging in third-party punishment, or social censure for not doing so (Barclay,

2006; Jordan et al., 2016; Jordan & Rand, 2017; Kurzban, DeScioli, & O'Brien, 2007). On this basis, we expected that adding another witness to the social situation we have been exploring in these experiments would motivate subjects' to punish on behalf of a stranger.

Experiment 5

Experiment 5 was run simultaneously with Experiment 4: subjects in Experiment 5 were recruited as individuals and run with a single friend pair from Experiment 4. However, the datasets are otherwise completely independent: the dataset for Experiment 4 contained only subjects who had brought a friend with them to the experiment, whereas the dataset for Experiment 5 contained only the subjects who came by themselves to the experiment. In Experiment 5, subjects either (a) were insulted by a stranger or (b) witnessed a stranger receive an insult from another stranger. As in Experiment 4, subjects viewed an introductory display that indicated whether any of the subjects were already acquainted with each other. In this way, we made subjects aware that two of the other three (fictive) subjects knew each other, and that the remaining subject was a "stranger" that did not know anyone. In both conditions, the insulter was the stranger (i.e., not part of the friend pair). We tested whether victims of insults would punish and report more anger than would witnesses of insulted strangers. We also tested whether witnesses of insulted strangers would report more empathy for the victim than for a neutral bystander.

Experiment 5 Method

Subjects

Subjects (N = 172; 101 female) were undergraduate students at the University of Miami who participated for partial course credit and $10. Before data analysis, we flagged all subjects

who during debriefing expressed suspicion that their interactions with other subjects had been staged (n = 18). These exclusions left our final sample at N = 154 (92 female).

Procedure

Subjects were seated at individual computers in private carrels and told they would be interacting with three other subjects—located either in the same room or in different rooms—over a computer network during their experimental sessions. To increase the believability of the interactions and promote subjects' memory of which subjects took what actions, subjects were identified to each other by name. If fewer than four subjects showed up for a session, the session was still run under the ruse that the other subjects were in other rooms; interactions with those subjects were fully staged. Subjects were randomly assigned to one of two conditions: (a) receive an insult or (b) witness a stranger receive an insult (see Figure 9). Hence, when subjects themselves were insulted, an "insulter" and two "non-insulter" subjects (who were friends with each other) were present. When subjects witnessed a stranger receive an insult, there was a "victim," an "insulter," and one "non-insulter" subject who was friends with the victim.

Experiment 5 Results

Means and standard deviations for all major variables appear in Table 14. Main analyses were conducted as in Experiments 2-4. See Tables 15 and 16 for full model results.

Manipulation check: Did subjects perceive the insulting review as unfair/inaccurate?

Subjects who received an insulting review of their own essay rated the review as less fair/accurate (M = 1.99, SD = 2.11) than the non-insulting (i.e. slightly positive) review of their essay (M = 7.49, SD = 1.40), pMCMC < .001, d_z = 2.39. Likewise, witnesses of insults also rated the insulting review as less fair/accurate (M = 3.09, SD = 2.21) than the non-insulting review of the same essay (M = 7.11, SD = 1.30), pMCMC <.001, d_z = 1.59. As in Experiments 2 and 4,

there was a significant review by target-of-insult interaction (see Table 15: Subjects who were personally insulted rated the insulting review as less fair/accurate than did subjects who read an insulting evaluation of another subject's essay, *p*MCMC < .001, *d* = .51. Recipients' and witnesses' ratings of the non-insulting review did not differ, *p*MCMC = .20, *d* = .28.

Did punishment vary as a function of whether the victim was the subject or a stranger?

First, we examined the between-subjects comparison of the amount of punishment that subjects administered to insulters as a function of whether they received the insult or witnessed a stranger receive the insult (see Figure 10). Victims of insults did not punish insulters *in absolute terms* more than witnesses of insults did, *p*MCMC = .385, *d* = .12.

Next, we examined the within-subjects comparison of punishment of the insulter *relative* to punishment of the other two people. Despite the lack of a between-subjects (victims-versus-witnesses) effect, there was a strong within-subjects effect for victims of insults: Victims of insults punished insulters (*M* = .32, *SD* = .94) more severely than they punished the two non-insulters (*M* = -.27, *SD* = .62; *M* = -.21, *SD* = .72), *p*MCMC < .001, d_z = .72 and *p*MCMC < .001, d_z = .63, respectively. This same within-subjects effect on punishment was weaker among witnesses of insults, though nonetheless statistically significant: Witnesses of insults also punished insulters (*M* = .21, *SD* = .85) more than non-insulters (*M* = -.07, *SD* = .80), *p*MCMC < .001, and victims (*M* = .03, *SD* = .80), *p*MCMC = .029, but the effect sizes for witnesses were only about half as large as they were for victims (d_z = .38 and .26, respectively).

To test whether this victim-versus-witness difference in relative punishment of insulters was significant, we again created punishment difference scores by subtracting the punishment each subject administered to each of the two non-insulters in each condition (i.e., victims and witnesses separately) from the punishment each subject administered to the insulter. Of the four

possible victim-versus-witness comparisons of these four within-subject difference scores, three were statistically significant (*p*MCMCs = .001 to .015) and one nearly so (*p*MCMC = .052)[8]. All of these differences pointed to the same conclusion: victims of insults punished their insulters, relative to non-insulters, to a greater extent than did mere witnesses of insults.

Did anger vary as a function of whether the victim was the subject or a stranger?

First, we evaluated the between-subjects comparison of the anger toward insulters that subjects reported as a function of whether they received the insult or witnessed a stranger receive the insult. Victims of insults reported more anger toward insulters (M = 1.54, SD = 1.69) than witnesses of insults did (M = 0.57, SD = .99), *p*MCMC < .001, d = .71.

Next, we evaluated the within-subjects comparisons of anger toward the insulter *relative* to anger toward the other two people. As Figure 3 shows, victims of insults reported more anger toward insulters than toward non-insulters (M = 0.28, SD = .82 and M = .27, SD = .68), *p*MCMC < .001, d_z = .78 and *p*MCMC < .001, d_z = .71, respectively). In contrast, witnesses *did not* report more anger toward insulters than toward non-insulters (M = 0.37, SD = .74), *p*MCMC = .142, d_z = .22 or toward victims (M = 0.37, SD = .85), *p*MCMC = .135, d_z = .20 (see Figure 10).

To test whether this victim-versus-witness difference in relative anger toward insulters was significant, we created difference scores for anger in the same way we did for punishment. Of the four possible victim-versus-witness comparisons of these four within-subject difference scores, all four were statistically significant (*p*MCMCs < .001). Thus, victims of insults reported more anger toward their insulters, relative to non-insulters, than did mere witnesses of insults.

[8] In the analyses with suspicious subjects included, two of these analyses remained significant (pMCMCs = .005 and .009) and two were not (pMCMCs = .096 and .155). All effects remained in the same direction.

Did differences in anger mediate differences in punishment for victims and witnesses of insulted strangers?

As noted above, both victims and witnesses of insults punished insulters to a greater extent than they punished non-insulters, but the relative difference was about twice as large for victims of insults. We were therefore interested in whether the between-condition (i.e., victim vs. witness of insulted stranger) differences in anger toward insulters versus non-insulters mediated the differences in the between-condition differences in punishment of insulters versus non-insulters. To explore this possibility, we conducted four mediation analyses (insulters vs. each of the two non-insulters in each condition) using the same method as we did for Experiments 2 and 4. In two of these four mediation models, the indirect effects were significant (b = .16, 95% CI : [.06, .31]; b = .12, 95% CI: [.003, .28]). In the other two mediation models, the indirect effects were not significant (b = .07, 95% CI : [-.04, .22]; b = .03, 95% CI : [-.10, .18]), though both were in the same direction as the significant effects. Thus, victims may have punished insulters more than non-insulters to a greater extent than witnesses did in part because victims were also angrier at insulters (relative to non-insulters) than witnesses were.

Did witnesses report empathy for victims?

Finally, we evaluated whether witnesses of insults experienced more empathy for victims than they did for non-insulters and insulters. Witnesses did not report more empathy toward victims (M = 1.76, SD = 1.15) than toward non-insulters (M = 1.71, SD = 1.24), pMCMC = .66, d_z = .05, but they did report more empathy toward victims than toward insulters (M = 1.38, SD = 1.19), pMCMC < .001, d_z = .39. Thus, if one thinks of subjects' empathy toward non-insulters as a control comparison, it seems reasonable to conjecture that witnesses of insults experienced reductions in empathy for insulters, but not increases in empathy for victims.

Experiment 5 Discussion

Experiment 5 yielded seven main results. First, targets of insults experienced a clear and focused motivation to punish: They punished insulters, relative to their punishment of non-insulters, about twice as harshly as did witnesses of insults. Second, in contrast to Experiments 2-4, witnesses of insults also punished insulters more than they punished non-insulters, which is consistent with the altruistic punishment hypothesis (Fehr & Fischbacher, 2004)—and other hypotheses (e.g., Barclay, 2006; Jordan et al., 2016; Kurzban et al., 2007)—although this effect was weaker than it was for victims of insults. Third, victims of insults experienced more anger toward insulters than they did toward non-insulters. Fourth, witnesses of insults *did not* experience more anger toward the insulter than they did toward non-insulters. Fifth, we found some evidence that victims punished insulters (relative to controls) more than witnesses did specifically because victims experienced more anger at insulters than witnesses did. Sixth, witnesses of insults did not report any more empathy for victims than they did for bystanders.

Aside from the altruistic punishment hypothesis, three possibilities come to mind for explaining the significant amount of third-party punishment we found in Experiment 5 (which we did not find in Experiments 2-4). First, it might reflect a Type 1 error (i.e., a false positive result), but more substantive explanations also seem tenable: The presence of a second witness, who subjects believed to be a friend of the victim, might have increased have increased the amount of social censure that subjects anticipated for not taking steps to appear concerned about upholding social norms (Barclay, 2006; Jordan et al., 2016; Kurzban et al., 2007; but see Balafoutas et al., 2014). Furthermore, the fact that friends have high levels of welfare interdependence means that the harm against the victim was also an indirect harm against the victim's friend (Pedersen et al., under review). Subjects' understanding of this social

consequence of friendship might have further increased their motivation to avoid social censure, thereby impelling them to punish the insulter.

Meta-Analytic Summary of Punishment and Anger from Experiments 2-5

The results for punishment and anger were relatively consistent across Experiments 2-5: In every experiment, victims of insults punished insulters significantly more than they punished non-insulters, and they reported more anger toward insulters than they did toward non-insulters. These results comport well with how subjects in Experiment 1 claimed they would respond when faced with a hypothetical vignette that depicts the same social situation. Conversely, three of four experiments yielded evidence that witnesses of insults directed toward strangers did *not* punish insulters any more than they punished non-insulters (contrary to subjects' claims in response Experiment 1's hypothetical vignette), though they did report more anger toward insulters than toward non-insulters in three of those four experiments (in line with the results of Experiment 1).

To summarize these results across experiments and compare them to results the results from Experiment 1, we combined data from Experiments 2-5 to conduct meta-analytic regressions for four outcomes: Victims' punishment of insulters; Victims' anger toward insulters; Witnesses' punishment of insulters on behalf of strangers; and Witnesses' anger toward insulters on behalf of strangers. For each subject, we created a punishment difference score by subtracting punishment toward the non-insulter from punishment toward the insulter using the same approach that we used to analyze the data from Experiments 2-5. We used a parallel approach to create anger difference scores for each subject. These difference scores reflect the surplus amounts of punishment and anger that subjects directed toward the insulter. Due to the different designs of the experiments, the only difference score for empathy we could create consistently for Experiments 2-5 (and combine for meta-analysis) would compare empathy for

victims to empathy for insulters, which would not reveal whether subjects were more empathic for victims, or less empathic for insulters, relative to non-insulters. Thus, we did not meta-analyze the empathy results. For each punishment and anger analysis, we ran an intercept-only, Bayesian linear mixed model predicting difference scores for either punishment or anger. Hence, each model simply estimated a posterior distribution for the mean difference score. We specified random intercepts for experiments to account for the nesting of observations within experiments and used non-informative priors for the models.

To further quantify the confidence one should place in the magnitude of the overall effects we found, we defined a so-called region of practical equivalence (ROPE; Kruschke, 2011) of the punishment and anger difference scores for each analysis. The ROPE concept enables a researcher to set a range of effect sizes that he or she would consider to be too small to be meaningfully different from zero. Using the subjective criteria we established (see below), we then determined the proportion of the posterior probability distribution that fell within the ROPE, enabling us to estimate the probability that punishment and anger toward insulters and non-insulters are *practically equivalent*. In addition to constructing ROPEs for each of the four meta-analytic regressions, we also constructed them for witnesses' punishment and anger on behalf of friends in Experiment 4 to assess the robustness of those findings.

We constructed two different ROPEs for each analysis. First, we considered effect sizes of Cohen's $d_z < .2$ to be practically equivalent to zero. We chose this effect size because it corresponds to the arbitrary threshold conventionally associated with "small" effects (Cohen, 1988). Second, we took a purely practical approach of determining what effect size would be detectable with 80% power given a reasonable time frame for data collection in a laboratory experiment. That is, we roughly estimated what effect size future research efforts might be well-

powered to detect, assuming similar data collection resources to our own. We collected usable data from a total of 676 subjects in Experiments 2-5 over the course of four semesters. With a sample size of 676 subjects, one would have 80% power to detect an effect size of approximately d_z = .11 in a dependent-samples t-test for each difference score. To construct the upper and lower boundaries of the ROPE for each analysis, we determined the raw difference score values that corresponded to +/- .2 SD or +/- .11 SD from 0 (Kruschke, 2011). See Figure 11 for posterior distributions and ROPEs of all analyses.

Hypothetical victims and witnesses: Punishment and anger

To put the meta-analytic results of the data from Experiments 2-5 in context, we first conducted ROPE analyses on the data for Experiment 1 (N = 456), in which subjects forecasted their responses to hypothetical vignettes in which they either were insulted directly or witnessed the insult of a stranger. Posterior distributions for Experiment 1 are displayed in light grey in Figure 11. For all four variables (anger and punishment for victims and witnesses), the ROPE analyses for $d_z < .2$ revealed that the probability of practical equivalence was 0 (i.e., no part of the posterior distributions fell within .2 SD of 0). These results indicate that we can be quite confident the true effect sizes for hypothetical victims and hypothetical witnesses' forecasts of their punishment and anger toward insulters (relative to their forecasts of their punishment and anger toward non-insulters), are larger than a $d_z = .20$.

Victims of insults: Punishment and anger

We combined the data for subjects who were victims (N = 196) in Experiments 2, 4, and 5 (in Experiment 3 subjects did not themselves receive insults). For each subject we created two difference scores by subtracting punishment (or anger) toward a non-insulter from punishment (or anger) toward the insulter. Experiments 4 and 5 had two possible "non-insulters": for

Experiment 4, the non-insulter was the stranger (rather than the subject's friend); for Experiment 5, we randomly selected the non-insulter from the two equivalent options.

Did victims of insults punish insulters more than they punished non-insulters? Yes. Overall, victims punished insulters (M = .33, SD = .94) significantly more than they punished non-insulters (M = -.26, SD = .72): model-estimated difference score = .58, 95% HDI: [.45, .71], pMCMC < .001, d_z = .65 (a "medium" effect size), just as subjects in Experiment 1 claimed they would. According to the ROPE analyses, the probability that punishment of insulters and punishment of non-insulters are practically equivalent for d_z < .2 is .00002, and the probability for d_z < .11 is 0 (i.e., no values in the posterior distribution fell within +/- .11 SD from 0; see Figure 11, top left panel). Thus, we can be very confident that the true effect size is greater than a Cohen's d_z of .2 and that future research efforts that involve similar amounts of data collection will be very well powered to find a significant effect (Kruschke, 2011).

Were victims of insults angrier toward insulters than they were toward non-insulters? Yes. Victims reported more anger toward insulters (M = 1.87, SD = 1.70) than toward non-insulters (M = .36, SD = .83): model-estimated difference score = 1.51, 95% HDI: [1.26, 1.75], pMCMC < .001, d_z = .87 (a "large" effect size), just as subjects in Experiment 1 claimed they would. According to the ROPE analyses, the probability that anger toward insulters and anger toward non-insulters are practically equivalent for both a d_z < .2 and d_z < .11 is .00004 (see Figure 11, top right panel). This analysis indicates that we can be confident that the true effect size is greater than a Cohen's d_z of .2, and that future research efforts with comparable numbers of subjects as analyzed here would be well-powered to find a statistically significant effect.

Witnesses of insults directed toward friends: Punishment and anger

Recall that in Experiment 4 we ran one condition in which subjects who were witnesses of insults directed toward their friends (N = 66). Subjects in this condition punished insulters more than they punished non-insulters, $d_z = .42$ (a "small" to "medium" effect). According to ROPE analyses for this condition, the probability that punishment toward insulters and toward non-insulters are practically equivalent for $d_z < .2$ is .039, and the probability for $d_z < .11$ is .007 (see Figure 11, middle left panel). This analysis indicates that we can be fairly confident that the true effect size is greater than a Cohen's d_z of .2, and also that future research with comparable numbers of subjects would be well-powered to find a statistically significant effect.

Similarly, in Experiment 4, subjects who were witnesses of insults that were directed toward their friends reported more anger toward insulters than toward non-insulters, $d_z = .60$ (a "medium" effect). According to the ROPE analyses, the probability that anger toward insulters and anger toward non-insulters are practically equivalent for $d_z < .2$ is .00008, and the probability for $d_z < .11$ is .00001 (see Figure 11, middle right panel). This analysis indicates that we can be confident that the true effect size is greater than a Cohen's d_z of .2 and that future research with comparable numbers of subjects would be well-powered to find a statistically significant effect.

Witnesses of insults directed toward strangers: Punishment and anger

Next, we combined the data for subjects who were witnesses of insults directed toward strangers ($N = 402$) in Experiments 2-5. For each subject we created two difference scores by subtracting punishment (or anger) toward the victim from punishment (or anger) toward the insulter. We chose to create the difference score in this way (rather than use the difference in punishment/anger between the insulter and a neutral bystander) because each of the four

experiments contained the comparison. (Experiments 2 and 3 did not have an uninvolved non-insulter as did Experiments 4 and 5.)

Did witnesses of insults punish insulters more than they punished victims? No. In contrast to subjects' forecasts in Experiment 1, witnesses of insults did not punish insulters (M = .04, SD = .84) more than they punished victims (M = -.003, SD = .81): model-estimated difference score = .05, 95% HDI: [-.03, .12], pMCMC = .227, d_z = .06. According to the ROPE analyses, the probability that punishment of insulters and punishment of victims are practically equivalent for $d_z < .2$ and $d_z < .11$ are .997 and .842, respectively (see Figure 11, bottom left panel). Thus, we can be reasonably confident that the true effect size is *smaller than* Cohen's d_z = .2, and that future research efforts of similar scope will be inadequately powered to find significant effects.

Were witnesses of insults angrier toward insulters than they were toward victims? Yes. Overall, similar to subjects' forecasts in Experiment 1, witnesses of insults reported more anger toward insulters (M = .72, SD = 1.05) than toward victims (M = .34, SD = .76): model-estimated difference score = .38, 95% HDI: [.27, .50], pMCMC < .001, d_z = .34 (a "small" effect size). This effect size was about half as large as both the effect size for subjects' forecasts from Experiment 1, d_z = .67, and the effect size for victims of insults in Experiments 2-5, d_z = .87). Even so, according to the ROPE analyses, the probability that anger toward insulters and anger toward victims are practically equivalent for $d_z < .2$ is .003, and the probability for $d_z < .11$ is 0 (see Figure 11, bottom right panel). These results indicate that we can be reasonably confident that the true effect size is greater than a Cohen's d_z of .2, and that future research with comparable numbers of subjects would be well-powered to find statistically significant differences in the amount of anger that witnesses of insults experience for insulters versus others.

Summary of Meta-Analytic Results

The meta-analytic results yielded robust evidence that victims of insults become angrier and more punitive toward people who have insulted them, or who have insulted their friends, than toward non-insulting bystanders. Further, the results indicate that people who have witnessed a stranger insult another stranger *do not* punish insulters more than they punish the victims themselves. To be sure, subjects in Experiment 5 did modestly punish people who had insulted strangers, but this finding may be the exception that proves the rule: The fact that the insulted stranger had a friend present in Experiment 5 might have elevated subjects' desire to avoid the social censure they might have faced by not punishing on the victim's behalf. In any case, our model suggests that there is a 99.7% chance that witnesses do not practically differ with an effect size $d_z \geq .2$ (and an 88.4% chance for $d_z \geq .11$) in the amount they punish insulters, relative to how much they punish victims, in the context of the social situations we created in these experiments. Witnesses do, however, report more anger toward insulters than they report toward victims of insults, although the effect size is lower than for victims ($d_z = .34$ vs. .87).

The effect sizes for witnesses in Experiments 2-5 who responded to an insult in real time were much smaller than for the hypothetical witnesses in Experiment 1 who responded to a vignette. Indeed, as the bottom panels in Figure 11 show, the posterior distributions for witnesses in Experiment 1 hardly overlap with those from Experiments 2-5. Hence, as in Pedersen et al. (2013), the discrepancy of our overall findings for witnesses from Experiments 2-5 from those in Experiment 1—and, more generally, the literature on third-party punishment—may be caused by a human desire *to be seen as punitive* rather than by a desire *to actually be punitive*. In contrast, hypothetical victims' emotional and behavioral responses to the insult depicted in the vignette of Experiment 1 were similar in size to those we obtained in Experiments 2-5 when subjects

responded in real time to the same insult. Thus, subjects who imagined themselves suffering an insult personally tended to provide affective and behavioral forecasts that corresponded reasonably well to how people actually respond in real time. Taken together, these results suggest that experiments that seek to measure punitive behavior using methods that create experimental demand for third-party punishment—as do all vignette studies and, plausibly, all previous third-party punishment experiments that explicitly prompt Adjusters to indicate how much third-party punishment they would like to enact—might systematically overestimate people's willingness to punish transgressors on behalf of strangers.

General Discussion

Revisiting the Altruistic Punishment Hypothesis

The hypothesis that third parties altruistically punish social norm-violators has played an important role in the past two decades of research on human cooperation (for reviews, see Fehr & Fischbacher, 2003; Guala, 2012; McCullough et al., 2013). However, some scientists are coming to doubt whether the empirical examples of third-party punishment in the extant literature should be classified as altruistic in an evolutionary sense (e.g., Krasnow et al., 2012; Krasnow, Delton, Cosmides, & Tooby, 2016; Pedersen et al., 2013; West et al., 2011). The results presented here give further cause for doubt: In experimental laboratory paradigms that were well-suited to elicit anger and retaliatory aggression, disinterested observers did not, on average, punish social norm-violators who had harmed strangers, even though they did experience a small amount of anger toward them. Nevertheless, people do appear to possess a robust anger-based inclination to punish people who have insulted them directly, along with a weaker anger-based inclination to punish people who have insulted their friends.

Our findings are consistent with ethnographic research on people from small-scale societies—whose way of life is the best available approximation of the social ecology in which ancestral humans evolved (Marlowe, 2005). In such societies, people rarely engage in costly punishment unless they or their kin directly suffer a serious harm (Boehm, 2008; Ericksen & Horton, 1992). Instead, people tend to simply ostracize exploitative individuals, which costs little and usually reforms uncooperative behavior. When third-party punishment on behalf of non-kin does occur, it is typically low-cost because it has community backing.

The present results corroborate Pedersen et al.'s (2013) conclusion that methodological artifacts such as experimental demand (Weber & Cook, 1972) might be largely responsible for the empirical results from the third-party punishment game that have led other scientists to conclude that humans possess a propensity for altruistically punishing norm-violators. That is, our results call into question whether it is a desire to punish that actually causes people's choices in the standard third-party punishment game, even though people's scores on this game are regularly assumed to be caused by such a desire. Instead, the score meaning (i.e., the construct that can be inferred to be the cause of Adjusters' observed behavior) may simply be a desire to *appear* morally motivated to punish norm-violators—a desire that is at times strong enough to impel people to pay for the opportunity to broadcast that appearance. The plausibility of this interpretation is bolstered by the strategic misreports of die rolls in Kriss et al.'s (2016) experiment on the third-party punishment game. In their experiment, in which subjects' die rolls determined whether Adjusters' costly commitments to punishing unfair Dictators would actually be enacted, Adjusters rolled the die *before* they learned whether Dictators had treated the Receiver unfairly. Most Adjusters who had committed to punishment chose to undo that commitment by lying about the outcome of the die roll (dictator game Receivers, on the other

hand, tended to strategically misreport the outcome of the die roll so that they could dishonestly *retain* their ability punish an unfair dictator). These results suggest that Adjusters knew that they did not actually want to punish on behalf of others at the moment of their punishment decisions, but committed to doing so, even though it cost them money, due to a desire *to seem committed to punishing violations of social norms*—not by a desire *to punish social norm violations*. Another possibility is that Adjusters who punish in the standard third-party punishment game do so because, in the constraints of the experimental design, it is the only way they can display their moral condemnation (Xiao & Houser, 2005)—which possibly functions to help form alliances with others (DeScioli & Kurzban, 2009). In our experiments, subjects were in a position to punish a disrespectful evaluation of an essay rather than an inequitable allocation of an economic windfall, which is the standard moral violation studied in the third-party punishment game. This methodological difference might cause some to doubt whether our findings are even germane to questions about the existence of altruistic third-party punishment. Such doubts would be misplaced. Third-party punishment has never been (nor could plausibly be) theorized as a phenomenon that is specific to breaches of fairness in sharing resources such as money. Researchers who are active in this research area regularly invoke the same psychological forces that motivate punishment behavior in the third party punishment game (and other economic games) to explain people's tendencies to sanction norm violations such as littering, failing to stand on the socially appropriate side of an escalator (e.g., Balafoutas & Nikiforakis, 2012; Balafoutas, Nikiforakis, & Rockenbach, 2014, 2016), and even desertion of one's comrades during warfare (Mathew & Boyd, 2011, 2014). If the same punitive desire that presumably causes people to punish stingy Dictators can be posited to be active in people's desires to chide litterers, chasten people who breach escalator etiquette, or administer whippings to battlefield

deserters, then surely it cannot be judged out of hand to be *not* active in laboratory subjects' angry and aggressive behavioral responses to unmerited insults to one's own (or others') academic opinions and writing. We cannot think of any argument that excludes our evidence from consideration on these questions that does not embrace either operationism (the long discredited thesis that theoretical constructs owe at least part of their nature to how they are measured; Green, 1992) or special pleading. In any case, the essence of validity lies in the successful manipulation and measurement of theory-relevant constructs, not in direct comparability with popular paradigms (Borsboom et al., 2004).

It is certainly possible that third parties decide how to punish inequity differently from how they decide how to respond to disrespectful remarks. It is also possible that third parties decide how to administer aversive sound blasts differently from how they decide how to deduct money. If these situations really do engage distinct psychological systems, then the value of our experiments here is in delimiting the boundary conditions in which altruistic punishment can be expected to obtain, which would be an important and meaningful advance in its own right. However, we think the implications of this work are more extensive inasmuch as they, when considered alongside the experimental evidence from Pedersen et al. (2013) and Kriss et al. (2016), corroborate the claim that third parties punish not out of concern for the victim, nor out of concern for defending social norms, but instead, out of concern for avoiding negative social evaluation (Kriss et al., 2016; Pedersen et al., 2013). Qualitatively similar findings using distinct methodologies bolster confidence that each operationalization of the construct is valid (Campbell & Fiske, 1959). Consequently, we believe a reader is justified in accepting the parsimonious conclusion that, across multiple types of moral violations, third parties do not possess a robust moralistic desire to altruistically punish norm-violators. Instead, we contend, the third-party

punishment on behalf of strangers that has been documented extensively in previous research has been motivated largely by self-presentation concerns, which can produce reputational benefits that increase the punisher's fitness (e.g., Barclay, 2006; Jordan et al., 2016; Jordan & Rand, 2017). If the behaviors are motivated by psychological systems that seek reputational benefits, then the behaviors themselves cannot be considered biologically altruistic.

Another potential objection to the present results is that the altruistic punishment hypothesis implies that people incur costs to punish norm violators on behalf of others, whereas our design allowed third parties to punish norm violators free of charge. To argue that our results do not undermine the altruistic punishment hypothesis on this basis would reveal a confusion between the proximate and evolutionary explanations of third-party punishment. Recall that the word "altruistic" in the altruistic punishment hypothesis is used in an evolutionary sense: The hypothesis posits that ancestral humans who possessed a moralistic desire to punish norm violators on behalf of others on average incurred a net fitness cost by acting on such a desire. The hypothesis does *not* imply that the moralistic desire itself is in any way defined by the instantaneous costs and benefits associated with implementing it; instead, the moralistic desire to punish that is posited to underlie altruistic punishment is wholly defined by two features: (a) it is activated upon the perception of a norm violation, and (b) it is satisfied when the norm violator has been punished. All we can infer about a desire that came to fixation via evolutionary altruism is that it is likely strong enough to have overcome countervailing, self-interested desires to avoid incurring personal costs.

It remains true, however, that a moralistic desire to engage in third-party punishment—like all desires that might motivate people to punish others (Guala, 2012; McCullough, et al., 2013)— is more likely to motivate punishment if the perceived costs are low. Thus, evaluating

whether third parties were willing to administer the cost-free punishment that our experiments afforded set up an especially low bar for the altruistic punishment hypothesis to clear. The fact that we, on average, did not observe third-party punishment in designs featuring cost-free punishment is consequently especially inconvenient for the altruistic punishment hypothesis. Furthermore, to argue that we might have observed more third-party punishment on behalf of strangers had it been costly would be to argue that there is some intrinsic feature of costs that *encourages* altruistic punishment. The only two plausible features for costs we can surmise in support of this argument are, first, that costs might increase the reputational value of third-party punishment—though, in such a case, punishment for reputational benefits would then not be classified as altruistic. Second, charging subjects money to enact a given behavior also makes clear that the behavior is of focal interest to the experimenters, potentially creating additional experimenter demand for punishment absent any willingness to do so otherwise.

If Third-Party Punishment is Not Altruistic, then Why *Do* People Punish?

If the punishment we observed in Experiments 2-4 was motivated by the judgment that insulters had violated a social norm (as the altruistic punishment hypothesis entails), then we would have observed punishment on behalf of the self, friends, and strangers alike. However, we observed punishment only on behalf of the self (i.e., second-party punishment) and friends (i.e., non-altruistic third-party punishment). This pattern suggests that the punishment was caused by a psychological concern that is present when people consider the welfare of themselves and their friends, but not when they consider the welfare of strangers.

One obvious possibility is that subjects valued themselves and their friends, but not the insulted strangers. Under this hypothesis, people who punished the insulter did so because they were angry about the *harm* the insults inflicted on themselves and their friends, not about the

violation of a moral norm per se (Batson, Kennedy, et al., 2007). An important implication is that the punishment was not morally motivated, but instead, was a response to personally meaningful interpersonal harm: Subjects valued themselves and so punished on their own behalf, and valued their friends (at some fraction of how much they valued themselves) and so punished on their behalf, but did not value the strangers enough to punish on their behalf. Subjects' modest amounts of third-party anger toward insulters is harder to explain. Perhaps participants did feel some moral anger, though not enough to motivate punishment. Or perhaps self-report measures of anger toward insulters in such situations can serve as a form of cheap talk in their own right. Adjudicating between these hypotheses will be important for resolving debates about whether moral anger exists at all (Batson, 2015).

Conclusion

The concept of altruistic punishment has become pivotal to many social scientists' understandings of human sociality. Indeed, so many scientists now take its existence for granted that most inquiry has moved beyond investigating whether a propensity for altruistic third-party punishment is even a real feature of human nature. Instead, most researchers now simply assert its reality and then seek to shed light on its manifestations in the field (Balafoutas & Nikiforakis, 2012; Balafoutas et al., 2014, 2016; Mathew & Boyd, 2011, 2014), its cross-cultural correlates (Henrich et al., 2006; Marlowe et al., 2008), its neural foundations (de Quervain et al., 2004; Singer et al., 2006; Strobel et al., 2011), and its basis in basic personality traits and individual differences (Crockett, Clark, Lieberman, Tabibnia, & Robbins, 2010; Johnson, Dawes, Fowler, McElreath, & Smirnov, 2009; Lotz, Baumert, Schlösser, Gresser, & Fetchenhauer, 2011).

Because the ontological reality of a human propensity for altruistic third-party punishment is so widely accepted, readers may be tempted to infer from our failures to find

strong evidence for its existence here or in our previous experiments (Pedersen et al., 2013) that something is wrong with our research rather than with the concept. One might, for instance, criticize our use of students as subjects, but even the very first studies on the third-party punishment game used students as subjects (Fehr and Fischbacher, 2004). In addition, our work was well controlled, high in experimental realism, and free of the methodological artifacts (viz., experimental demand for punishment) that characterize most of experimental work on this topic (e.g., Bernhard et al., 2006; Fehr & Fischbacher, 2004; Henrich et al., 2006; Marlowe et al., 2008). And as we explained above, our focus on aggressive responses to insults is not an inferential weakness, but instead, a virtue.

Altruistic punishment has been invoked to explain a variety of social phenomena from the laboratory and the field, and has proven a useful in-principle solution to theoretical puzzles arising from game theory and the field of social evolution (e.g., Bowles & Gintis, 2004; Boyd et al., 2003; Fehr & Fischbacher, 2003; Fehr & Gächter, 2002). However, such theoretical work has demonstrated only that altruistic punishment is a plausible *potential* explanation for social behavior that *could* have evolved under a variety of conditions. Too little research on altruistic punishment has sought to confirm its ontological reality—the essence of validity (Borsboom et al., 2004)—which was the motivation behind the present work. It is our hope that our results will simulate further discussion not only about the reality of altruistic punishment, but also, about how researchers can increase the validity of experiments so that experimental psychologists and other social scientists may draw trustier conclusions about human cooperation.

Context of the Research

Humans sometimes experience (and act upon) a desire to punish individuals who have harmed them, but the popular scientific claim that people also possess a desire to punish norm violators on behalf of others has been more controversial. Some researchers have found, for instance, that punishment of norm violators on behalf of strangers in economic games may be caused by experimenter demand effects. We carried out the present studies to test whether people possess an intrinsic desire to punish on behalf of mistreated strangers using experiments that do not involve economic games. The results confirmed that people possess (a) a relatively strong desire to punish people who have harmed them directly; (b) a robust albeit weaker desire to punish people who have harmed their friends; and (c) little or no desire to punish strangers who have harmed other strangers (in the absence of experimental features that may artificially promote punishment). These findings are consistent with the hypothesis that people desire to punish on behalf of valued others. Future directions include testing this welfare interdependence hypothesis in non-laboratory settings, as well as determining whether people desire to punish on behalf of friends as an end in itself or, instead, to avoid censure from their friends.

References

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351-371.

Anderson, C. A., & Bushman, B. J. (1997). External validity of "trivial" experiments: The case of laboratory aggression. *Review of General Psychology, 1*, 19-41.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390-1396.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412.

Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: a natural field experiment. *European Economic Review, 56*(8), 1773-1785.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature communications, 7*.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*, 325-344.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics, 11*(2), 122-133.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3).

Bastian, B., Jetten, J., & Fasoli, F. (2011). Cleansing the Soul by Hurting the Flesh The Guilt-Reducing Effect of Pain. *Psychological Science, 22*(3), 334-335.

Batson, C. D. (2015). *What's wrong with morality?: a social-psychological perspective*: Oxford University Press.

Batson, C. D., Eklund, J. H., Chermok, V. L., Hoyt, J. L., & Ortiz, B. G. (2007). An additional antecedent of empathic concern: valuing the welfare of the person in need. *Journal of Personality and Social Psychology, 93*(1), 65.

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

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36*(1), 59-122.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior, 10*, 122-142.

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

Boehm, C. (1987). *Blood revenge: The enactment and management of conflict in Montenegro and other tribal societies* (2nd ed.). Philadelphia: University of Pennsylvania Press.

Boehm, C. (2008). Purposive social selection and the evolution of human altruism. *Cross-Cultural Research, 42*, 319-352.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127-135.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology, 65*(1), 17-28.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*, 3531-3535.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology, 13*, 171-195.

Burnham, T. C., & Johnson, D. D. P. (2005). The biological and evolutionary logic of human cooperation. *Analyse & Kritik, 27*(2), 113-135.

Bushman, B. J., & Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: Does self-love or self-hate lead to violence? *Journal of Personality and Social Psychology, 75*, 219-229.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.

Chagnon, N., & Bugos, P. (Eds.). (1979). *Kin selection and conflict: An analysis of a Yanomamö ax fight*. North Scituate: Duxbury.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology, 39*, 268-277.

Clutton-Brock, T. H. (1989). Mammalian mating systems. *Proceedings of the Royal Society B:, 236*, 339-372.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society,* 1431-1451.

Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., & Robbins, T. W. (2010). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion, 10,* 855-862.

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305,* 1254-1258.

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, 108,* 13335-13340.

DeScioli, P., & Kurzban, R. (2009). Mysteries of morality. *Cognition, 112*(2), 281-299.

Ericksen, K. P., & Horton, H. (1992). "Blood feuds": Cross-cultural variations in kin group vengeance. *Behavior Science Research, 26,* 57-85.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785-791.

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

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*(1), 1-25.

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

Frey, K. S., Pearson, C. R., & Cohen, D. (2015). Revenge is seductive, if not sweet: Why friends matter for prevention efforts. *Journal of Applied Developmental Psychology, 37,* 25-35.

Giancola, P. R., & Chermack, S. T. (1998). Construct validity of laboratory aggression paradigms: A response to tedeschi and quigley (1996). *Aggression and Violent Behavior, 3*(3), 237-253.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*(2), 169-179.

Green, C. D. (1992). Of immortal mythological beasts: Operationism in psychology. *Theory & Psychology, 2*(3), 291-320.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35*(01), 1-15.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software, 33*(2), 1-22.

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology, 69*, 339-348.

Halevy, N., & Halali, E. (2015). Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. Proceedings of the National Academy of Sciences, 112(22), 6937-6942.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I, II. *Journal of Theoretical Biology, 7*, 1-52.

Harmon-Jones, E., & Sigelman, J. (2001). State anger and prefrontal brain activity: Evidence that insult-related relative left prefrontal activity is associated with experienced anger and aggression. *Journal of Personality and Social Psychology, 80*, 797–803.

Hendricks, A. (2012). SoPHIE - Software Platform for Human Interaction Experiments. . *University of Osnabrueck, working paper*.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & al., e. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*, 795-855.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science, 327*, 1480-1484.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Ziker, J. (2006). Costly punishment across human societies. *Science, 312*, 1767-1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*, 1362-1367.

Huitsing, G., & Veenstra, R. (2012). Bullying in classrooms: Participant roles from a social network perspective. *Aggressive Behavior, 38*(6), 494-509.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters, 102*(3), 192-194.

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

Jordan, J. J., McAuliffe, K., & Rand, D. (2015). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 1-23.

Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology, 421*, 189-202.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business*, S285-S300.

Karmali, F., Kawakami, K., & Page-Gould, E. (2017). He Said What? Physiological and Cognitive Responses to Imagining and Witnessing Outgroup Racism. *Journal of Experimental Psychology. General*.

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science, 323*(5911), 276-278.

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What Are Punishment and Reputation for? *PLoS One, 7*(9), e45662.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*.

Kriss, P. H., Weber, R. A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: on the robustness of costly punishment. *Journal of Economic Behavior & Organization*.

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science, 6*(3), 299-312.

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.

Leliveld, M. C., Dijk, E., & Beest, I. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology, 42*(2), 135-140.

Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature, 445*, 727-731.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*(3), 482-493.

Lotz, S., Baumert, A., Schlösser, T., Gresser, F., & Fetchenhauer, D. (2011). Individual differences in third‐party interventions: How justice sensitivity shapes altruistic punishment. *Negotiation and Conflict Management Research, 4*(4), 297-313.

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology, 47*(2), 477-480.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*: Routledge.

Marlowe, F. W. (2005). Hunter-gatherers and human evolution. *Evolutionary Anthropology: Issues, News, and Reviews, 14*(2), 54-67.

Marlowe, F. W. (2009). Hadza cooperation. *Human Nature, 20*(4), 417-430.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., . . . Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society of London, Series B--Biological Sciences, 275*, 587-590.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences, 108*(28), 11375-11380.

Mathew, S., & Boyd, R. (2014). The cost of cowardice: punitive sentiments towards free riders in Turkana raids. *Evolution and Human Behavior, 35*(1), 58-64.

McAuliffe, W. H. (2017). *The Psychology of Common Knowledge Explains the Appearance of Altruistic and Moral Motivation.* University of Miami, Open Access Theses. Retrieved from http://scholarlyrepository.miami.edu/oa_theses/682

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences, 36*(01), 1-15.

McCullough, M. E., Pedersen, E. J., Tabak, B. A., & Carter, E. C. (2014). Conciliatory gestures promote forgiveness and reduce anger in humans. *Proceedings of the National Academy of Sciences*.

Nelissen, R. M. A. (2012). Guilt-induced self-punishment as a sign of remorse. *Social Psychological and Personality Science, 3*(2), 139-144.

Nelissen, R. M. A. (2014). Relational utility as a moderator of guilt in social interactions. *Journal of Personality and Social Psychology, 106*(2), 257.

Nelissen, R. M. A., & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a Dobby Effect. *Emotion, 9*(1), 118.

O'Mara, E. M., Jackson, L. E., Batson, C. D., & Gaertner, L. (2011). Will moral outrage stand up? Distinguishing among emotional reactions to a moral violation. *European Journal of Social Psychology, 41*, 173-179.

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences, 280*(1758).

Pedersen, E. J., McAuliffe, W. H. B., & McCullough, M. E. (under review). Welfare interdependence, anger, and the regulation of third-party punishment: A net cost model.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary psychology and criminal justice: A recalibrational theory of punishment and reconciliation. In H. Høgh-Oleson (Ed.), *Human morality and sociality: Evolutionary and comparative perspectives* (pp. 72-131). New York: Palgrave MacMillan.

Phillips, S., & Cooney, M. (2005). Aiding peace, abetting violence: Third parties and the management of conflict. *American Sociological Review, 70*, 334-354.

Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour, 70*(4), 901-908.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136-168).

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439*(7075), 466-469.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*(1), 83-117.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *NeuroImage, 54*(1), 671-680.

Thomas, K. A., De Freitas, J., DeScioli, P., & Pinker, S. (2016). Recursive mentalizing and common knowledge in the bystander effect. *Journal of Experimental Psychology: General, 145*(5), 621.

Tooby, J., & Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions* (3rd ed., pp. 114-137). New York: Guilford.

Tooby, J., Cosmides, L., Sell, A. N., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In A. J. Elliott (Ed.), *Handbook of approach and avoidance motivation* (pp. 251-271). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Turillo, C. J., Folger, R., Lavelle, J. J., Umphress, E. E., & Gee, J. O. (2002). Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organizational Behavior and Human Decision Processes, 89*(1), 839-865.

Twenge, J. M., Baumeister, R. F., Tice, D. M., & Stucke, T. S. (2001). If you can't join them, beat them: effects of social exclusion on aggressive behavior. *Journal of Personality and Social Psychology, 81*(6), 1058.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77*(4), 273.

West, S. A., El Mouden, C., & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior, 32*, 231-262.

Winking, J., & Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America, 102*(20), 7398-7401.

Table 1.

Reliabilities (Cronbach's alpha) for major study variables, Experiments 1-5

Experiment	Target	Punishment	Anger	Empathy	Fairness/Accuracy
1	Insulter	.78	.93	.89	NA
	Non-Insulter	.73	.90	.90	
2	Insulter	.71	.91	.89	Insult: .97
	Non-Insulter	.69	.83	.87	Neutral: .82
3	Insulter	.58	.92	.82	Insult: .87
	Non-Insulter	.63	.86	.83	Neutral: .88
4	Insulter	.63	.94	.77	Insult: .88
	Non-Insulter	.52	.89	.83	Neutral: .69
	Friend	.70	.89	.79	
5	Insulter	.57	.94	.84	Insult: .87
	Non-Insulter 1	.37	.94	.82	Neutral: .80
	Non-Insulter 2	.55	.84	.84	

Table 2.

Experiment 1 descriptive statistics for major study variables

		Forecasted Punishment		Forecasted Anger		Forecasted Empathy	
Condition	Target	M	SD	M	SD	M	SD
Victim	Insulter	0.34	1.05	1.98	1.36	0.82	0.99
n = 228	Non-Insulter	-0.36	0.81	0.64	1.03	1.47	1.16
Witness	Insulter	0.25	0.80	1.45	1.23	0.83	0.91
n = 228	Non-Insulter	-0.24	0.72	0.48	0.84	2.27	1.21

Note. Condition: Victim = subject imagined receiving an insult; Witness = subject imagined witnessing a stranger insult another stranger.

Table 3.

Experiment 1 linear mixed model results predicting forecasted punishment and forecasted anger

Parameter	Forecasted Punishment			Forecasted Anger		
	b	95% HDI	*p*MCMC	b	95% HDI	*p*MCMC
Intercept	0.34	[0.23, 0.45]	< .001	1.98	[1.84, 2.13]	< .001
Non-Insulter	-0.71	[-0.85, -0.57]	< .001	-1.34	[-1.55, -1.14]	< .001
Condition: Witness	-0.09	[-0.25, 0.06]	0.266	-0.53	[-0.74, -0.32]	< .001
Non-Insulter*Condition	0.21	[0.02, 0.42]	0.036	0.36	[0.07, 0.66]	0.017

Note. Results of two linear mixed models, one predicting forecasted punishment and one predicting forecasted anger. Predictors were dummy coded, and the in these models the intercept refers to the insulter in the victim condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Table 4.

Experiment 2 descriptive statistics for major study variables, collapsed across levels of the empathy manipulation

		Punishment		Anger		Empathy	
Condition	Target	M	SD	M	SD	M	SD
Victim	Insulter	0.35	0.96	2.16	1.59	0.84	1.09
n = 55	Non-Insulter	-0.20	0.80	0.54	0.92	2.12	1.58
Witness	Insulter	-0.01	0.86	0.74	0.92	1.36	1.32
n = 62	Non-Insulter	-0.12	0.82	0.36	0.77	2.53	1.47

Note. Condition: victim = subject received an insult; witness = subject witnessed a stranger insult another stranger.

Table 5.

Experiment 2 linear mixed model predicting empathy (manipulation check)

Parameter	b	95% HDI	*p*MCMC
Intercept	2.59	[2.12, 3.08]	< .001
Insulter	-1.30	[-1.98, -0.63]	< .001
Condition: No empathy	-0.13	[-0.81, 0.56]	0.714
Condition: Victim	0.02	[-0.66, 0.70]	0.949
Insulter*No Empathy	0.27	[-0.68, 1.23]	0.571
Insulter*Victim	-0.33	[-1.28, 0.63]	0.493
No Empathy*Victim	-1.00	[-2.00, 0.04]	0.054
Insulter*No Empathy*Victim	0.53	[-0.83, 1.99]	0.461

Note. Results of a linear mixed model predicting empathy. Predictors were dummy coded, and the intercept refers to the non-insulter when the subject was a witness in the empathy condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Table 6.

Experiment 2 linear mixed model results predicting fairness/accuracy of the reviews (manipulation check)

Parameter	b	95% HDI	*p*MCMC
Intercept	3.44	[2.87, 4.00]	< .001
Neutral review	3.78	[2.97, 4.59]	< .001
Condition: Victim	-1.05	[-1.89, -0.22]	0.014
Neutral review*Victim	1.22	[0.05, 2.41]	0.043

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were dummy coded, and the intercept refers to the insulting review when the subject was a witness.

Table 7.

Experiment 2 linear mixed model results predicting punishment and anger, collapsed across levels of the empathy manipulation

	Punishment			Anger		
Parameter	b	95% HDI	*p*MCMC	b	95% HDI	*p*MCMC
Intercept	0.35	[0.11, 0.58]	0.004	2.16	[1.87, 2.45]	< .001
Non-Insulter	-0.54	[-0.80, -0.29]	< .001	-1.62	[-2.02, -1.20]	< .001
Condition: Witness	-0.36	[-0.69, -0.05]	0.030	-1.42	[-1.82, -1.02]	< .001
Non-Insulter*Condition	0.43	[0.08, 0.78]	0.015	1.24	[0.66, 1.79]	< .001

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were dummy coded, and the in these models the intercept refers to the insulter in the victim condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Table 8.

Experiment 3 descriptive statistics for major study variables

Future Interaction	Partner Generosity	Target	Punishment		Anger		Empathy	
			M	SD	M	SD	M	SD
No n = 105	Fair n = 43	Insulter	-0.13	0.92	1.02	1.35	1.75	1.31
		Non-Insulter	-0.02	0.87	0.46	0.88	2.21	1.35
	Generous n = 29	Insulter	-0.21	0.71	1.09	1.32	1.25	1.14
		Non-Insulter	-0.22	0.64	0.40	0.88	2.18	1.33
	Very Generous n = 33	Insulter	0.22	0.77	0.82	1.06	1.58	1.09
		Non-Insulter	0.12	0.82	0.40	0.91	2.04	1.32
Yes n = 101	Fair n = 26	Insulter	-0.06	0.88	0.63	0.82	1.63	1.24
		Non-Insulter	0.10	1.14	0.46	0.83	2.01	1.28
	Generous n = 37	Insulter	0.01	0.99	0.85	1.14	1.71	1.31
		Non-Insulter	-0.07	0.76	0.38	0.65	2.15	1.36
	Very Generous n = 38	Insulter	0.15	0.80	0.48	0.81	0.89	0.91
		Non-Insulter	0.08	0.78	0.22	0.64	1.54	1.26

Table 9.

Experiment 3 linear mixed model predicting fairness/accuracy of the reviews (manipulation check)

Parameter	b	95% HDI	*p*MCMC
Intercept	5.36	[5.17, 5.56]	< .001
Insult	-2.13	[-2.32, -1.93]	< .001
Future: yes	-0.17	[-0.37, 0.02]	0.086
Generosity: fair	-0.07	[-0.35, 0.20]	0.604
Generosity: generous	-0.07	[-0.34, 0.21]	0.635
Insult*Future: yes	0.10	[-0.09, 0.29]	0.310
Insult*Generosity: fair	-0.13	[-0.40, 0.15]	0.359
Insult*Generosity: generous	-0.03	[-0.31, 0.24]	0.818
Future: yes*Generosity: fair	-0.12	[-0.39, 0.16]	0.398
Future: yes*Generosity: generous	0.21	[-0.07, 0.48]	0.141
Insult*Future: yes*Generosity: fair	-0.08	[-0.35, 0.20]	0.585
Insult*Future: yes*Generosity: generous	-0.02	[-0.29, 0.26]	0.901

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were effect coded, and the in these models the intercept refers to the grand mean. “Future” refers to the prospect of future interaction manipulation (yes or no); “Generosity” refers to the partner generosity manipulation (fair, generous, very generous). Reference categories are non-insulters, Future: no, Generosity: very generous.

Table 10.

Experiment 3 linear mixed model results predicting punishment and anger

	Punishment			Anger		
	b	95% HDI	*p*MCMC	b	95% HDI	*p*MCMC
Intercept	0.00	[-0.11, 0.10]	0.957	0.60	[0.50, 0.71]	< .001
Insulter	0.00	[-0.05, 0.06]	0.969	0.21	[0.13, 0.30]	< .001
Future: yes	0.04	[-0.07, 0.15]	0.490	-0.10	[-0.20, 0.01]	0.070
Generosity: fair	-0.02	[-0.18, 0.12]	0.747	0.04	[-0.11, 0.19]	0.598
Generosity: generous	-0.12	[-0.27, 0.03]	0.115	0.08	[-0.07, 0.22]	0.302
Insulter*Future: yes	0.00	[-0.05, 0.05]	0.982	-0.06	[-0.15, 0.02]	0.149
Insulter*Generosity: fair	-0.06	[-0.14, 0.01]	0.105	-0.03	[-0.15, 0.09]	0.592
Insulter*Generosity: generous	0.02	[-0.05, 0.10]	0.541	0.08	[-0.04, 0.20]	0.212
Future: yes*Generosity: fair	0.01	[-0.14, 0.16]	0.896	0.00	[-0.15, 0.15]	0.997
Future: yes*Generosity: generous	0.06	[-0.10, 0.21]	0.471	0.03	[-0.12, 0.18]	0.668
Insulter*Future: yes*Generosity: fair	-0.01	[-0.09, 0.06]	0.764	-0.03	[-0.16, 0.09]	0.592
Insulter*Future: yes*Generosity: generous	0.02	[-0.06, 0.10]	0.662	0.01	[-0.12, 0.13]	0.891

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were effect coded, and the in these models the intercept refers to the grand mean. "Future" refers to the prospect of future interaction manipulation (yes or no); "Generosity" refers to the partner generosity manipulation (fair, generous, very generous). Reference categories are non-insulters, Future: no, Generosity: very generous.

Table 11.

Experiment 4 descriptive statistics for major study variables

Condition	Target	Punishment		Anger		Empathy	
		M	SD	M	SD	M	SD
Victim n = 69	Insulter	0.33	0.95	2.00	1.73	0.78	1.01
	Non-Insulter	-0.28	0.76	0.36	0.86	1.40	1.34
	Friend	-0.18	0.97	0.15	0.67	2.14	1.54
Witness: Friend n = 66	Insulter	0.08	0.74	1.04	1.26	1.13	1.06
	Non-Insulter	-0.20	0.68	0.29	0.58	1.63	1.24
	Friend	0.01	1.01	0.13	0.35	2.23	1.31
Witness: Stranger n = 64	Insulter	0.02	0.76	0.60	1.02	1.24	1.01
	Non-Insulter	0.07	0.68	0.20	0.53	1.49	1.32
	Friend	0.18	0.92	0.14	0.53	2.12	1.42

Table 12.

Experiment 4 linear mixed model predicting fairness/accuracy of the reviews (manipulation check)

Parameter	b	95% HDI	*p*MCMC
Intercept	3.54	[3.07, 3.99]	< .001
Neutral review	4.17	[3.52, 4.82]	< .001
Victim	-1.18	[-1.82, -0.55]	< .001
Witness: Friend	-1.35	[-2.01, -0.72]	< .001
Neutral*Victim	1.20	[0.29, 2.08]	0.008
Neutral*Witness: Friend	1.20	[0.31, 2.13]	0.010

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were dummy coded, and the intercept refers to the insulting review when the subject witnessed a stranger receive an insult.

Table 13.

Experiment 4 linear mixed model results predicting punishment and anger

	Punishment			Anger		
	b	95% HDI	*p*MCMC	b	95% HDI	*p*MCMC
(Intercept)	0.33	[0.13, 0.53]	0.002	2.00	[1.77, 2.22]	< .001
Non-Insulter	-0.60	[-0.82, -0.39]	< .001	-1.64	[-1.93, -1.35]	< .001
Friend	-0.51	[-0.72, -0.30]	< .001	-1.85	[-2.14, -1.56]	< .001
Witness: Friend	-0.24	[-0.53, 0.04]	0.095	-0.96	[-1.29, -0.65]	< .001
Witness: Stranger	-0.31	[-0.59, -.02]	0.034	-1.40	[-1.72, -1.08]	< .001
Non-Insulter*Witness: Friend	0.32	[0.01, 0.62]	0.038	0.89	[0.49, 1.32]	< .001
Friend*Witness: Friend	0.44	[0.13, 0.74]	0.006	0.94	[0.52, 1.35]	< .001
Non-Insulter*Witness: Stranger	0.66	[0.35, 0.96]	< .001	1.24	[0.84, 1.67]	< .001
Friend*Witness: Stranger	0.67	[0.36, 0.97]	< .001	1.38	[0.96, 1.80]	< .001

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were dummy coded, and the in these models the intercept refers to the insulter in the victim condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Table 14.

Experiment 5 descriptive statistics for major study variables

Condition	Target	Punishment		Anger		Empathy	
		M	SD	M	SD	M	SD
Victim n = 75	Insulter	0.32	0.94	1.54	1.69	0.76	1.11
	Non-Insulter 1	-0.27	0.62	0.28	0.82	1.18	1.28
	Non-Insulter 2	-0.21	0.72	0.27	0.68	1.43	1.38
Witness n = 79	Insulter	0.21	0.85	0.57	0.99	1.38	1.19
	Non-Insulter 1	0.03	0.80	0.37	0.85	1.76	1.15
	Non-Insulter 2	-0.07	0.80	0.37	0.74	1.71	1.24

Table 15.

Experiment 5 linear mixed model results predicting fairness/accuracy of the reviews (manipulation check)

Parameter	b	95% HDI	*p*MCMC
Intercept	3.09	[2.69, 3.49]	< .001
Neutral review	4.02	[3.45, 4.58]	< .001
Condition: Victim	-1.10	[-1.66, -0.51]	< .001
Neutral review*Victim	1.47	[0.66, 2.27]	0.001

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were dummy coded, and the intercept refers to the insulting review when the subject was a witness.

Table 16.

Experiment 5 linear mixed model results predicting punishment and anger

	b	95% HDI	*p*MCMC	b	95% CI	*p*MCMC
Intercept	0.32	[0.13, 0.49]	0.001	1.54	[1.32, 1.78]	< .001
Non-Insulter 1	-0.59	[-0.76, -0.42]	< .001	-1.27	[-1.54, -0.99]	< .001
Non-Insulter 2	-0.53	[-0.69, -0.36]	< .001	-1.28	[-1.55, -1.00]	< .001
Condition: Witness	-0.11	[-0.36, 0.15]	0.399	-0.97	[-1.29, -0.64]	< .001
Non-Insulter 1*Condition: Witness	0.41	[0.19, 0.65]	< .001	1.06	[0.68, 1.44]	< .001
Non-Insulter 2*Condition: Witness	0.25	[0.02, 0.48]	0.037	1.08	[0.70, 1.46]	< .001

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were dummy coded, and in these models the intercept refers to the insulter in the victim condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

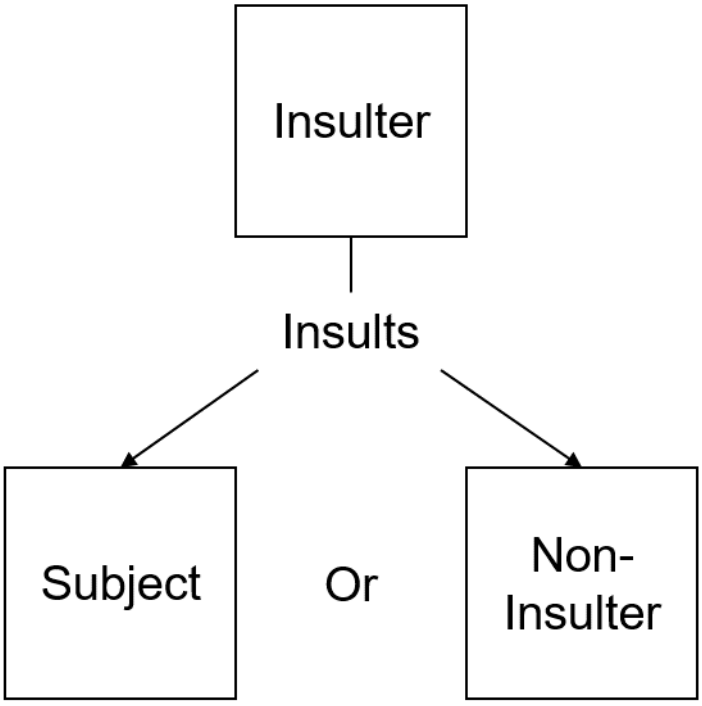

Figure 1. Schematic of roles in Experiments 1 and 2. In Experiment 1, subjects imagined either receiveing an insult from the insulter or witnessing the insulter insult the non-insulter. In Experiment 2, conducted in the laboratory in a situation high in experimental realism, subjects either received an actual insult from the insulter or witnessed the insulter insult the non-insulter. Additionally, in Experiment 2, subjects read either an empathy-inducing essay from the non-insulter or a neutral essay from the non-insulter prior to the insult.

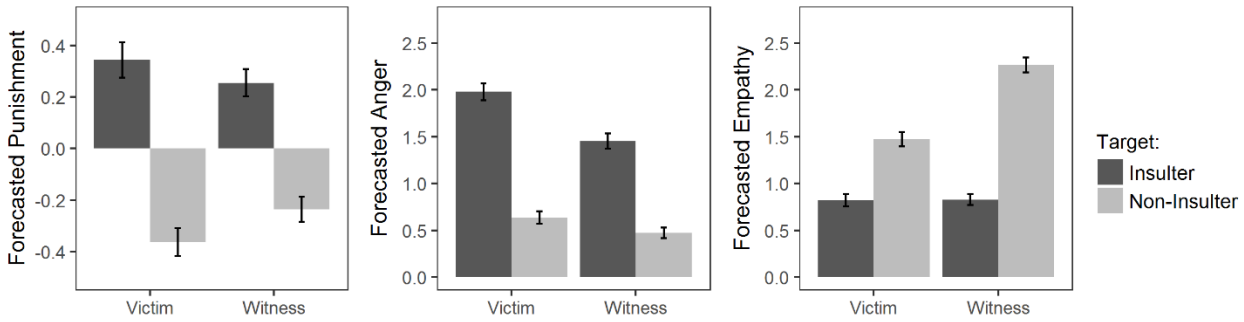

Figure 2. Experiment 1 means of forecasted punishment, forecasted anger, and forecasted empathy toward insulters and non-insulters as a function of whether subects imagined themselves receiving an insult (Victim) or witnessing a stranger receive an insult (Witness). Error bars = +/- 1 SE.

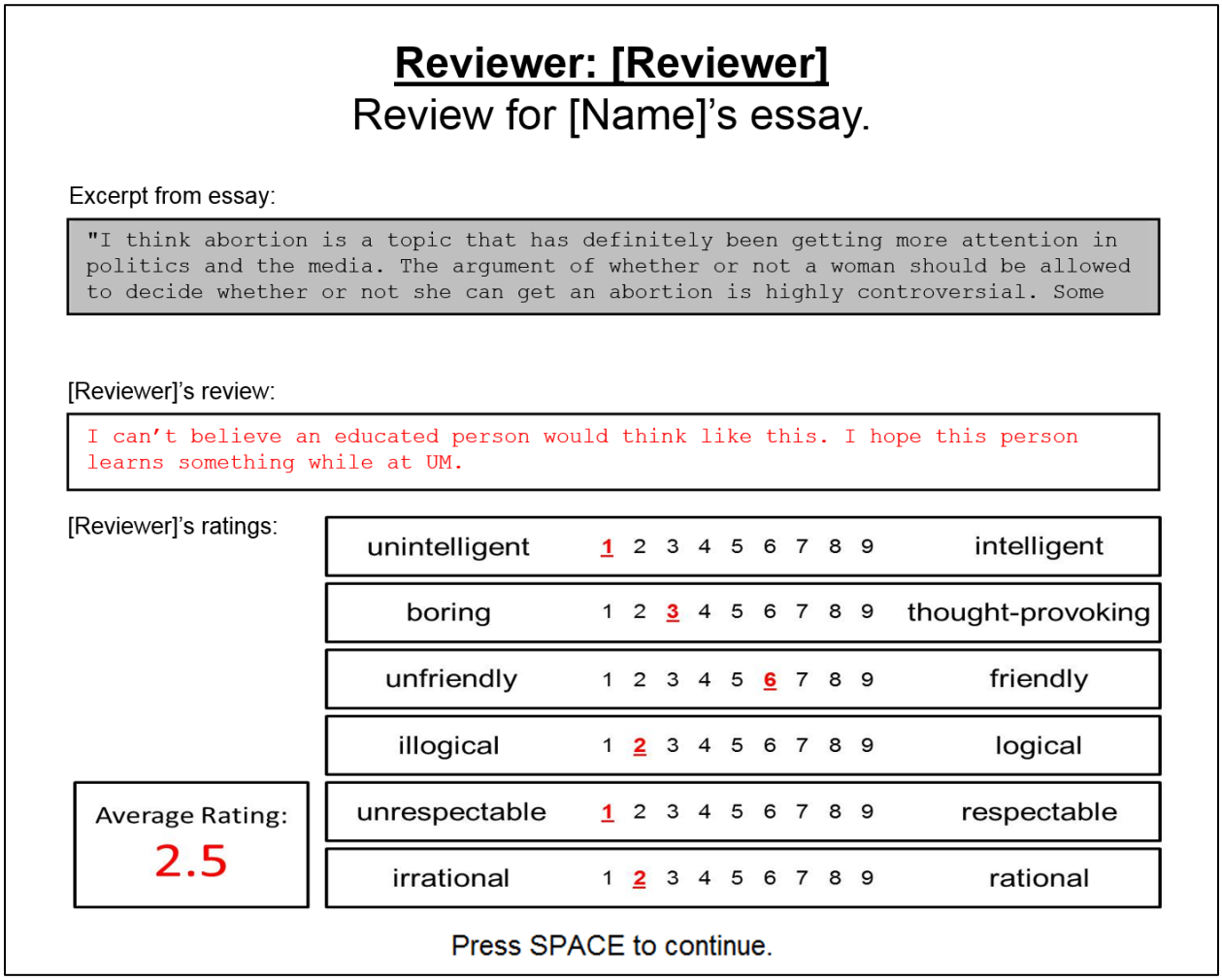

Figure 3. Screenshot of review of insulting essay evaluation.

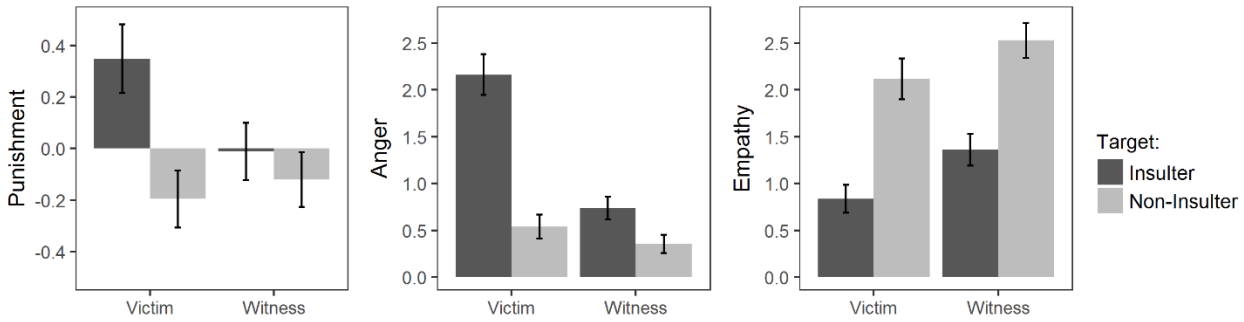

Figure 4. Experiment 2 means of punishment, anger, and empathy toward insulters and non-insulters as a function of whether subects received an insult (Victim) or witnessed a stranger receive an insult (Witness), collapsed across the empathy manipulation. Error bars = +/- 1 SE.

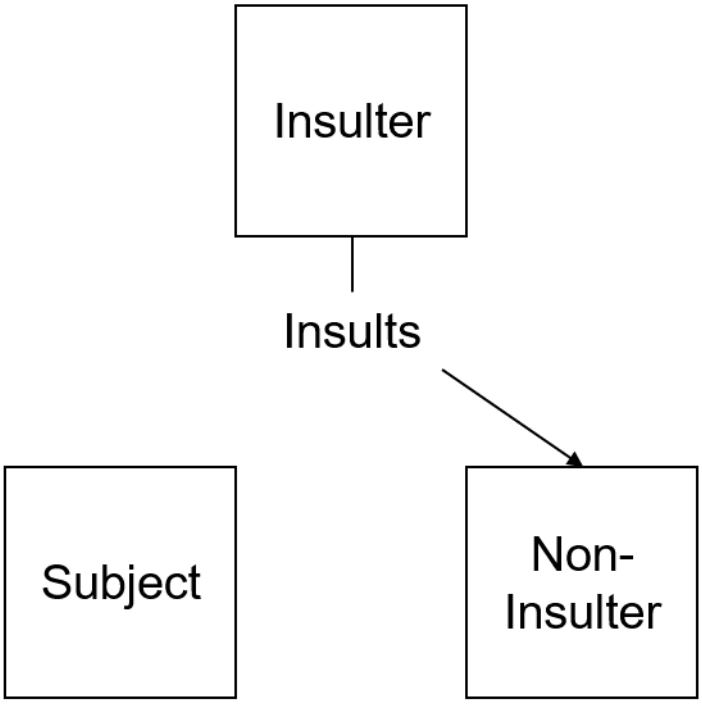

Figure 5. Schematic of roles in Experiment 3. Subjects always witnessed the insulter insult the non-insulter and assigned to one of six conditions in a 3 (partner generosity: low, medium, high) by 2 (prospect of future cooperative interaction: low, high) between-subjects design.

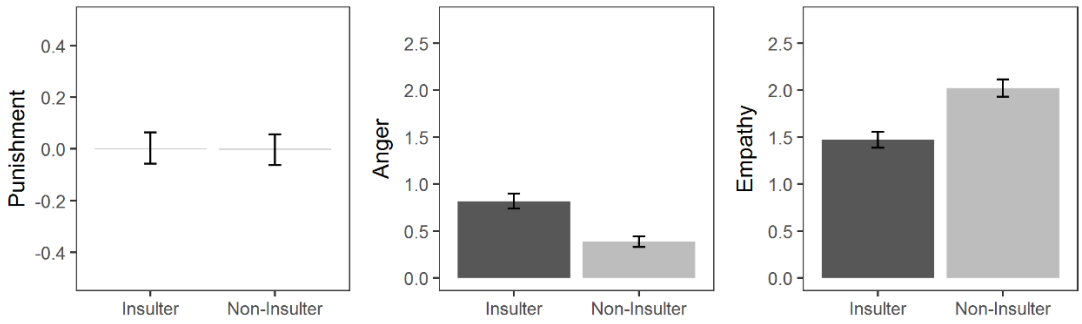

Figure 6. Experiment 3 means of punishment, anger, and empathy toward insulter and non-insulters, collapsed across prospect of future interaction and partner generosity manipulations. All subjects in Experiment 3 witnessed the insulter insult the non-insulter.

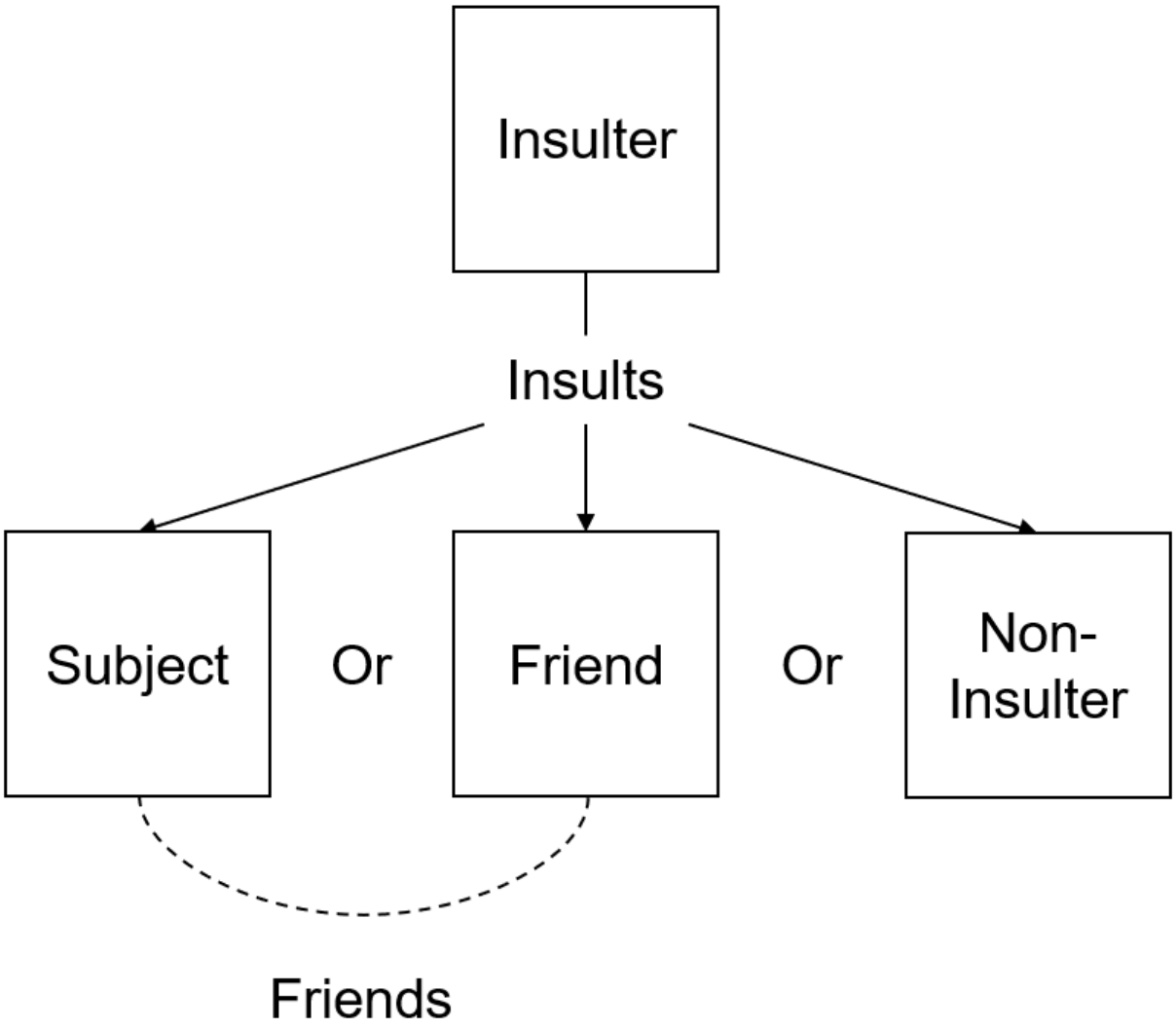

Figure 7. Schematic of roles in Experiment 4. Subjects either received an insult from the insulter, witnessed their friend receive an insult from the insulter, or witnessed the non-insulter receive an insult from the insulter.

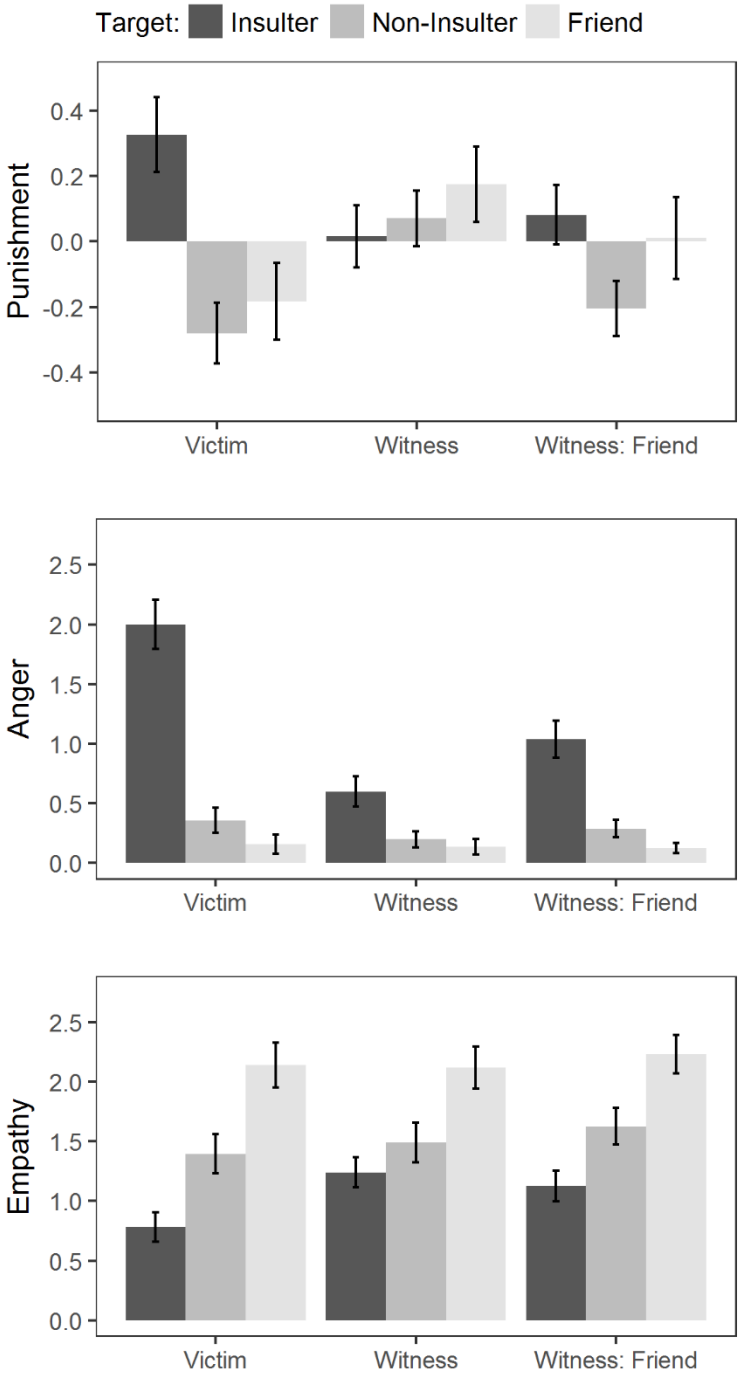

Figure 8. Experiment 4 means of punishment, anger, and empathy as function of whether subects received an insult (Victim), witnessed a stranger receive an insult (Witness), or witnessed their friend receive an insult (Witness: Friend). Error bars = +/- 1 SE.

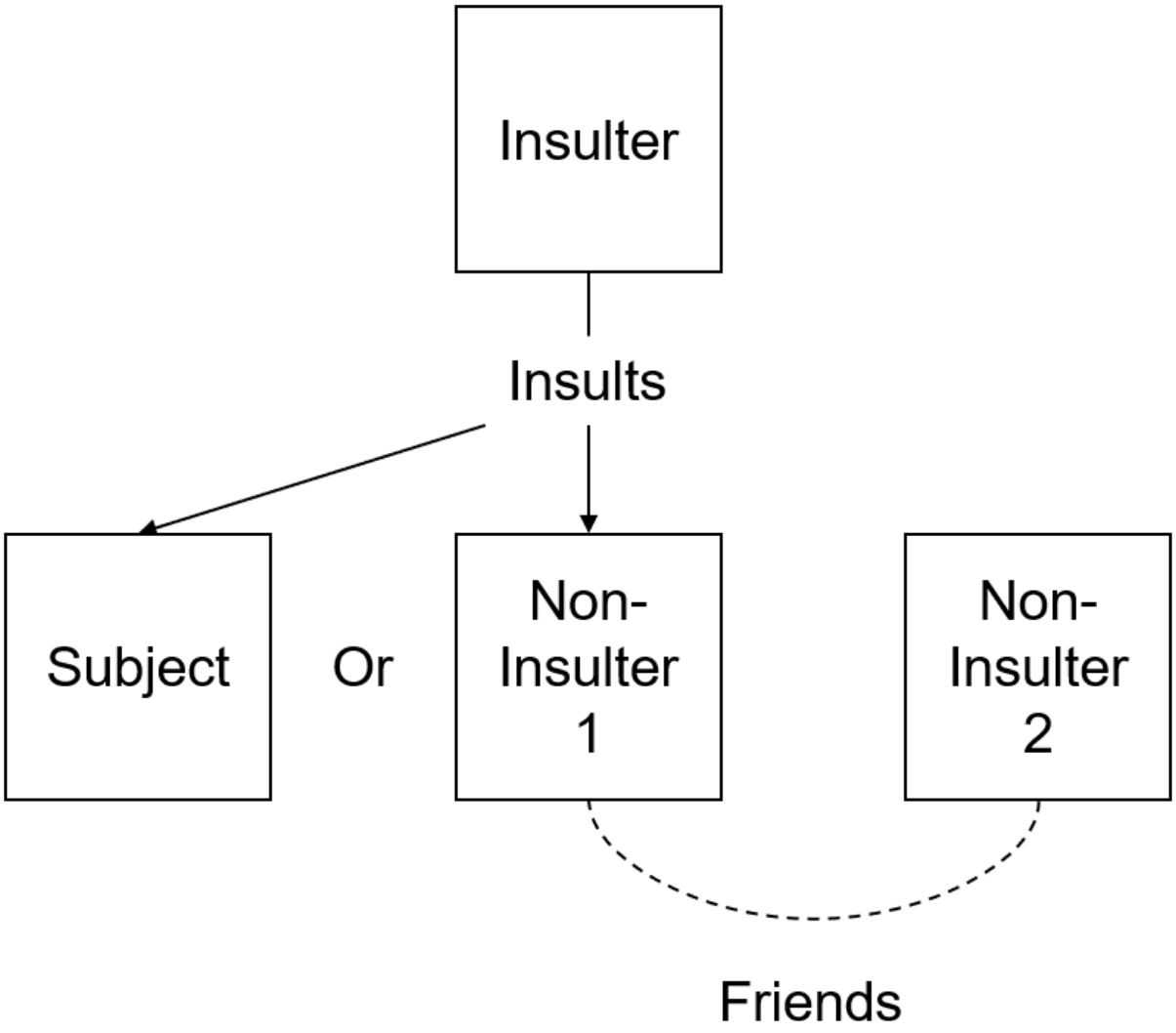

Figure 9. Schematic of roles in Experiment 5. Subjects either received an insult from the insulter or witnessed a non-insulter receive an insult from the insulter. Subjects were aware that the two non-insulters were friends. The non-insulters in Experiment 5 were the subjects in Experiment 4 and had no knowledge that they were "participants" in Experiment 5. Thus, the datasets analyzed for each Experiment were entirely independent.

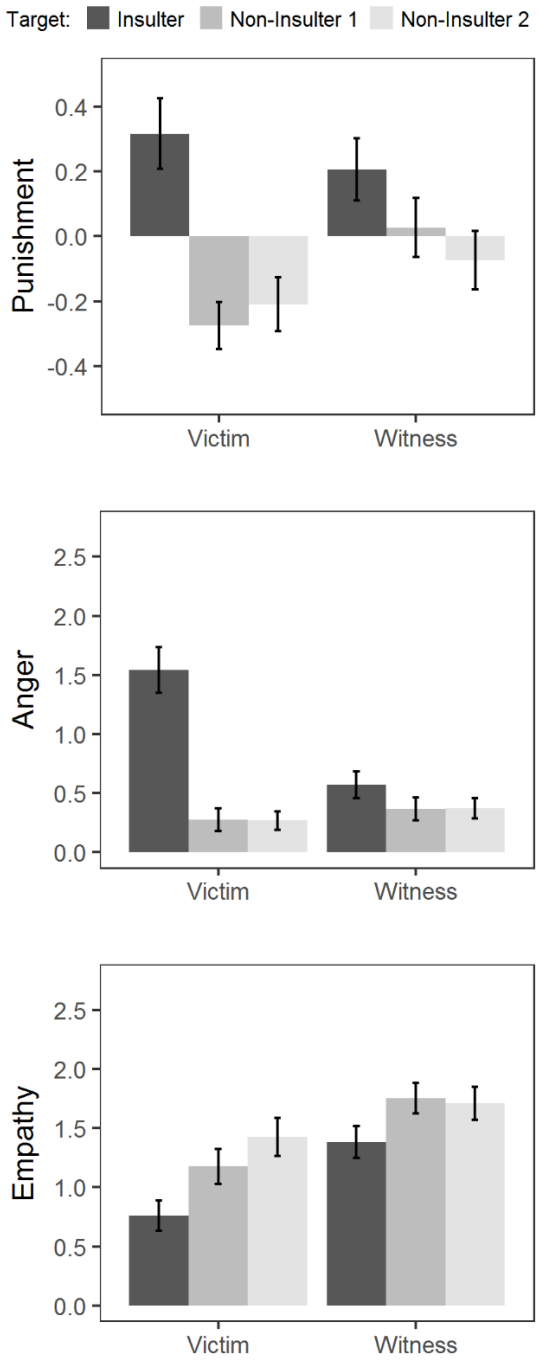

Figure 10. Experiment 5 means of punishment, anger, and empathy as a function of whether subects received an insult (Victim) or witnessed a stranger receive an insult (Witness). In the Witness condition, Non-Insulter 1 was the recipient of the insult and Non-Insulter 2 was an uninvolved bystander. Error bars = +/- 1 SE.

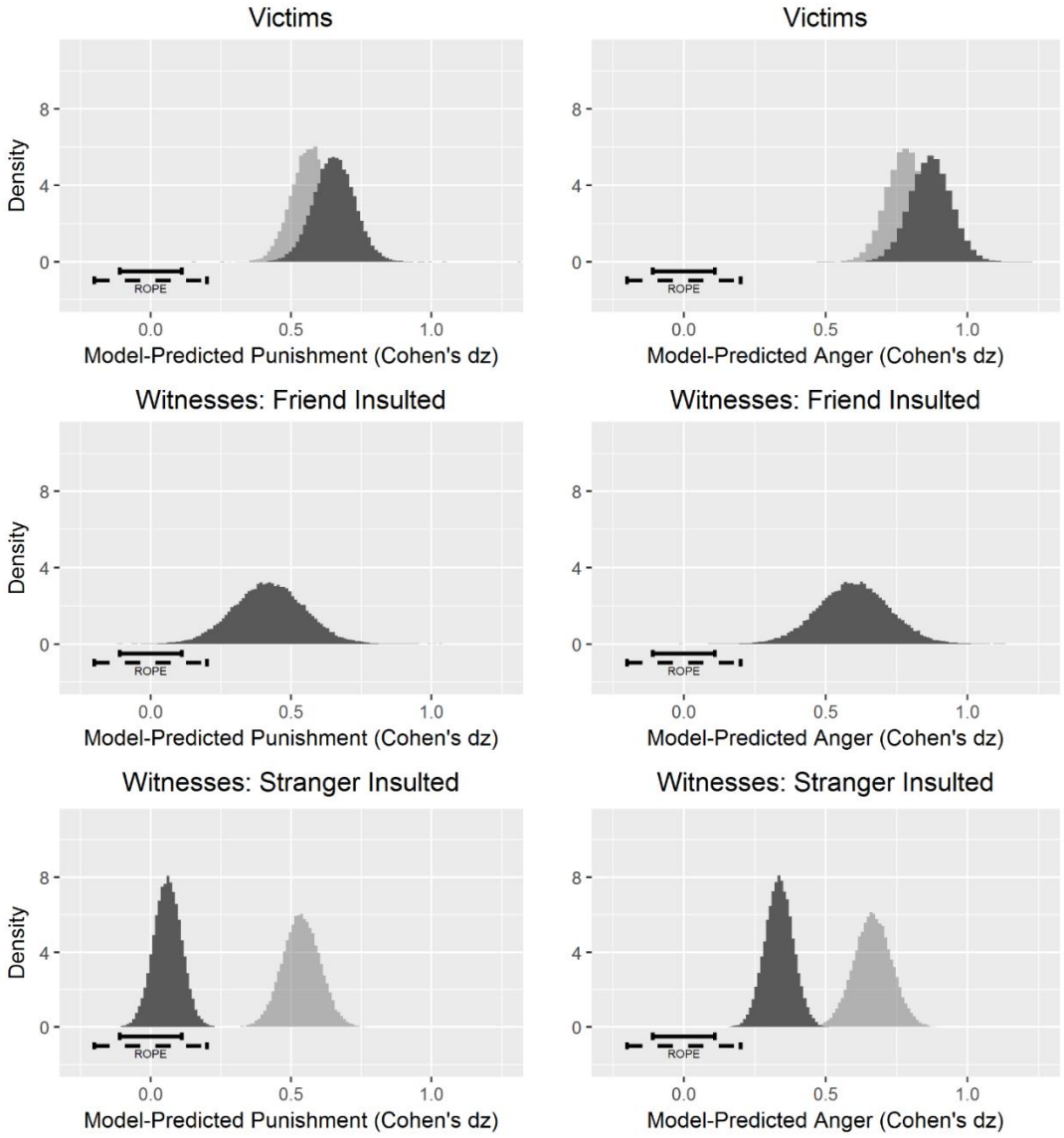

Figure 11. Posterior distributions of Cohen's d_z effect sizes for meta-analytic models of difference scores. Light grey distributions are derived from hypothetical forecasts from Experiment 1 (Victims N = 228; Witnesses: Stranger Insulted N = 228) and dark grey distributions are derived from real responses from Experiments 2-5 (Victims N = 196; Witnesses: Friend Insulted N = 66; Witnesses: Stranger Insulted N = 402). The solid line corresponds to a region of practical equivalence (ROPE) of d_z = 0 +/- .11 and the dashed line corresponds to a ROPE of d_z = 0 +/- .20.

Supplemental Material

This supplemental material contains descriptives and results tables from intent to treat analyses conducted on the full data sets (i.e., including all suspicious subjects). These tables are replications of those that appear in the main text, with suspicious subjects included. All analyses in the main text were conducted and reported prior to analyzing any data from suspicious subjects, and the results reported in the main text do not qualitatively differ from those in the intent to treat analyses unless noted by a footnote in the main text.

Additionally, Appendix A contains the script for Experiment 1.

Contents:

Table S1.

Experiment 2 descriptive statistics for major study variables, collapsed across empathy manipulation, including subjects flagged for suspicion

		Punishment		Anger		Empathy	
Condition	Target	M	SD	M	SD	M	SD
Victim	Insulter	0.31	0.99	2.09	1.53	0.82	1.08
n = 76	Non-Insulter	-0.22	0.73	0.48	0.86	2.16	1.58
Witness	Insulter	0.02	0.89	0.80	0.97	1.41	1.34
n = 71	Non-Insulter	-0.11	0.83	0.36	0.73	2.50	1.42

Note. Condition: victim = subject received an insult; witness = subject witnessed a stranger insult another stranger.

Table S2.

Experiment 2 linear mixed model predicting empathy (manipulation check), including subjects flagged for suspicion

Parameter	b	95% HDI	*p*MCMC
Intercept	2.59	[2.14, 3.03]	< .001
Insulter	-1.24	[-1.88, -0.61]	< .001
Condition: No empathy	-0.19	[-0.85, 0.44]	0.556
Condition: Victim	0.09	[-0.54, 0.71]	0.771
Insulter*No Empathy	0.31	[-0.58, 1.23]	0.505
Insulter*Victim	-0.44	[-1.32, 0.44]	0.323
No Empathy*Victim	-0.85	[-1.76, 0.01]	0.061
Insulter*No Empathy*Victim	0.38	[-0.89, 1.60]	0.454

Note. Results of a linear mixed model predicting empathy. Predictors were dummy coded, and the intercept refers to the non-insulter when the subject was a witness in the empathy condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Table S3.

Experiment 2 linear mixed model results predicting fairness/accuracy of the reviews (manipulation check), including subjects flagged for suspicion

Parameter	b	95% HDI	*p*MCMC
Intercept	3.36	[2.86, 3.87]	< .001
Neutral review	3.95	[3.25, 4.67]	< .001
Condition: Victim	-1.08	[-1.81, -0.40]	0.003
Neutral review*Victim	1.20	[0.20, 2.18]	0.019

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were dummy coded, and the intercept refers to the insulting review when the subject was a witness.

Table S4.

Experiment 2 linear mixed model results predicting punishment and anger, collapsed across empathy manipulation, including subjects flagged for suspicion

	Punishment			Anger		
Parameter	b	95% HDI	*p*MCMC	b	95% HDI	*p*MCMC
Intercept	0.30	[0.10, 0.50]	0.003	2.09	[1.84, 2.33]	< .001
Non-Insulter	-0.53	[-0.74, -0.31]	< .001	-1.60	[-1.96, -1.26]	< .001
Condition: Witness	-0.28	[-0.57, 0.00]	0.057	-1.29	[-1.63, -0.92]	< .001
Non-Insulter*Condition	0.39	[0.08, 0.71]	0.014	1.16	[0.67, 1.66]	< .001

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were dummy coded, and the in these models the intercept refers to the insulter in the victim condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Table S5.

Experiment 3 descriptive statistics for major study variables, including subjects flagged for suspicion

Future Interaction	Partner Generosity	Target	Punishment		Anger		Empathy	
			M	SD	M	SD	M	SD
No n = 125	Fair n = 48	Insulter	-0.19	0.92	0.95	1.29	1.67	1.28
		Non-Insulter	-0.12	0.92	0.42	0.84	2.07	1.37
	Generous n = 37	Insulter	-0.15	0.76	1.10	1.31	1.16	1.06
		Non-Insulter	-0.22	0.68	0.41	0.83	2.32	1.40
	Very Generous n = 40	Insulter	0.21	0.76	0.95	1.14	1.38	1.13
		Non-Insulter	0.03	0.83	0.41	0.91	2.04	1.38
Yes n = 125	Fair n = 37	Insulter	0.06	0.87	0.81	1.00	1.50	1.16
		Non-Insulter	0.19	1.07	0.41	0.79	2.04	1.22
	Generous n = 45	Insulter	0.04	0.91	0.89	1.21	1.59	1.26
		Non-Insulter	-0.12	0.73	0.35	0.61	2.23	1.40
	Very Generous n = 43	Insulter	0.22	0.84	0.63	0.99	0.92	0.91
		Non-Insulter	0.08	0.75	0.22	0.61	1.59	1.36

Table S6.

Experiment 3 linear mixed model predicting fairness/accuracy of the reviews (manipulation check), including subjects flagged for suspicion

Parameter	b	95% HDI	*p*MCMC
Intercept	5.23	[5.06, 5.41]	< .001
Insult	-2.23	[-2.41, -2.05]	< .001
Future: yes	-0.17	[-0.36, 0.00]	0.058
Generosity: fair	-0.06	[-0.32, 0.19]	0.637
Generosity: generous	-0.06	[-0.32, 0.19]	0.642
Insult*Future: yes	0.04	[-0.13, 0.22]	0.620
Insult*Generosity: fair	-0.19	[-0.44, 0.06]	0.125
Insult*Generosity: generous	0.05	[-0.21, 0.29]	0.699
Future: yes*Generosity: fair	-0.08	[-0.33, 0.17]	0.521
Future: yes*Generosity: generous	0.18	[-0.07, 0.44]	0.151
Insult*Future: yes*Generosity: fair	-0.12	[-0.37, 0.13]	0.338
Insult*Future: yes*Generosity: generous	0.01	[-0.25, 0.26]	0.958

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were effect coded, and the in these models the intercept refers to the grand mean. “Future” refers to the prospect of future interaction manipulation (yes or no); “Generosity” refers to the partner generosity manipulation (fair, generous, very generous). Reference categories are non-insulters, Future: no, Generosity: very generous.

Table S7.

Experiment 3 linear mixed model results predicting punishment and anger, including subjects flagged for suspicion

	Punishment			Anger		
	b	95% HDI	*p*MCMC	b	95% HDI	*p*MCMC
Intercept	0.00	[-0.09, 0.10]	0.970	0.63	[0.54, 0.72]	< .001
Insulter	0.03	[-0.02, 0.08]	0.228	0.26	[0.17, 0.35]	< .001
Future: yes	0.08	[-0.02, 0.17]	0.117	-0.08	[-0.17, 0.01]	0.089
Generosity: fair	-0.02	[-0.15, 0.12]	0.821	0.02	[-0.11, 0.15]	0.771
Generosity: generous	-0.12	[-0.25, 0.02]	0.090	0.06	[-0.07, 0.19]	0.374
Insulter*Future: yes	0.00	[-0.05, 0.05]	0.968	-0.03	[-0.12, 0.05]	0.432
Insulter*Generosity: fair	-0.08	[-0.15, -0.01]	0.020	-0.03	[-0.14, 0.10]	0.662
Insulter*Generosity: generous	0.03	[-0.04, 0.10]	0.383	0.05	[-0.07, 0.17]	0.432
Future: yes*Generosity: fair	0.07	[-0.07, 0.20]	0.338	0.04	[-0.09, 0.17]	0.537
Future: yes*Generosity: generous	0.00	[-0.14, 0.13]	0.953	0.01	[-0.12, 0.14]	0.875
Insulter*Future: yes*Generosity: fair	-0.01	[-0.08, 0.05]	0.713	0.00	[-0.11, 0.13]	0.951
Insulter*Future: yes*Generosity: generous	0.02	[-0.05, 0.09]	0.517	0.00	[-0.12, 0.12]	0.945

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were effect coded, and the in these models the intercept refers to the grand mean. "Future" refers to the prospect of future interaction manipulation (yes or no); "Generosity" refers to the partner generosity manipulation (fair, generous, very generous). Reference categories are non-insulters, Future: no, Generosity: very generous.

Table S8.

Experiment 4 descriptive statistics for major study variables, including subjects flagged for suspicion

Condition	Target	Punishment		Anger		Empathy	
		M	SD	M	SD	M	SD
Victim n = 77	Insulter	0.27	0.97	1.91	1.70	0.79	1.01
	Non-Insulter	-0.30	0.75	0.33	0.82	1.40	1.31
	Friend	-0.21	0.95	0.14	0.63	2.15	1.55
Witness: Friend n = 75	Insulter	0.11	0.74	1.07	1.29	1.09	1.07
	Non-Insulter	-0.21	0.68	0.28	0.57	1.55	1.23
	Friend	0.04	1.02	0.12	0.34	2.18	1.37
Witness: Stranger n = 70	Insulter	0.04	0.75	0.60	0.99	1.21	1.01
	Non-Insulter	0.08	0.67	0.19	0.51	1.53	1.33
	Friend	0.21	0.92	0.15	0.55	2.11	1.44

Table S9.

Experiment 4 linear mixed model predicting fairness/accuracy of the reviews (manipulation check), including subjects flagged for suspicion

Parameter	b	95% HDI	*p*MCMC
Intercept	3.54	[3.08, 3.99]	< .001
Neutral review	4.17	[3.53, 4.82]	< .001
Victim	-1.19	[-1.81, -0.55]	< .001
Witness: Friend	-1.33	[-1.98, -0.70]	< .001
Neutral*Victim	1.19	[0.33, 2.10]	0.008
Neutral*Witness: Friend	1.20	[0.31, 2.12]	0.010

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were dummy coded, and the intercept refers to the insulting review when the subject witnessed a stranger receive an insult.

Table S10.

Experiment 4 linear mixed model results predicting punishment and anger, including subjects flagged for suspicion

	Punishment			Anger		
	B	95% HDI	*p*MCMC	b	95% HDI	*p*MCMC
(Intercept)	0. 27	[0. 08, 0.46]	0.005	1.91	[1.70, 2.11]	< .001
Non-Insulter	-0.57	[-0.77, -0.37]	< .001	-1.57	[-1.85, -1.31]	< .001
Friend	-0.47	[-0.67, -0.28]	< .001	-1.77	[-2.05, -1.51]	< .001
Witness: Friend	-0.16	[-0.43, 0.11]	0.242	-0.84	[-1.14, -0.55]	< .001
Witness: Stranger	-0. 23	[-0.50, .04]	0.099	-1.30	[-1.61, -1.01]	< .001
Non-Insulter*Witness: Friend	0. 25	[-0.03, 0.53]	0.079	0.79	[0.40, 1.17]	< .001
Friend*Witness: Friend	0. 40	[0.12, 0.68]	0.006	0.82	[0.44, 1.21]	< .001
Non-Insulter*Witness: Stranger	0. 61	[0.31, 0.89]	< .001	1.16	[0.77, 1.55]	< .001
Friend*Witness: Stranger	0.64	[0.36, 0.93]	< .001	1.32	[0.93, 1.71]	< .001

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were dummy coded, and the in these models the intercept refers to the insulter in the victim condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Table S11.

Experiment 5 descriptive statistics for major study variables, including subjects flagged for suspicion

Condition	Target	Punishment		Anger		Empathy	
		M	SD	M	SD	M	SD
Victim n = 88	Insulter	0.33	0.90	1.61	1.72	0.74	1.08
	Non-Insulter 1	-0.21	0.67	0.33	0.93	1.16	1.23
	Non-Insulter 2	-0.17	0.73	0.23	0.63	1.42	1.34
Witness n = 84	Insulter	0.19	0.86	0.62	1.05	1.36	1.22
	Non-Insulter 1	0.00	0.78	0.41	0.97	1.76	1.18
	Non-Insulter 2	-0.13	0.83	0.35	0.72	1.70	1.25

Table S12.

Experiment 5 linear mixed model results predicting fairness/accuracy of the reviews (manipulation check) , including subjects flagged for suspicion

Parameter	b	95% HDI	*p*MCMC
Intercept	3.11	[2.71, 3.51]	< .001
Neutral review	4.02	[3.45, 4.57]	< .001
Condition: Victim	-1.11	[-1.68, -0.55]	< .001
Neutral review*Victim	1.48	[0.66, 2.26]	0.001

Note. Results of a linear mixed model predicting fairness/accuracy. Predictors were dummy coded, and the intercept refers to the insulting review when the subject was a witness.

Table S13.

Experiment 5 linear mixed model results predicting punishment and anger

	b	95% HDI	*p*MCMC	b	l.95..CI	*p*MCMC
Intercept	0.33	[0.16, 0.50]	<.001	1.61	[1.39, 1.83]	< .001
Non-Insulter 1	-0.54	[-0.70, -0.38]	< .001	-1.28	[-1.56, -1.00]	< .001
Non-Insulter 2	-0.50	[-0.67, -0.35]	< .001	-1.38	[-1.66, -1.11]	< .001
Condition: Witness	-0.14	[-0.38, 0.10]	0.258	-0.99	[-1.31, -0.67]	< .001
Non-Insulter 1*Condition: Witness	0.34	[0.12, 0.57]	.002	1.07	[0.68, 1.47]	< .001
Non-Insulter 2*Condition: Witness	0.18	[-0.05, 0.40]	0.120	1.11	[0.70, 1.49]	< .001

Note. Results of two linear mixed models, one predicting punishment and one predicting anger. Predictors were dummy coded, and the in these models the intercept refers to the insulter in the victim condition. As needed, we recoded the dummy codes and re-ran the models to obtain *p*-values for specific contrasts reported in the text.

Appendix A

How Would You Respond?

For this study, we would like for you to imagine yourself in a particular scenario in our laboratory. We will call this scenario "The Situation" throughout this study. We will ask you to complete a series of questions regarding how you think you would think, feel, and act in "The Situation." Please read carefully. Later, we will ask you some basic questions about what you read to ensure that you paid attention.

Part 1:

The Situation

Imagine that you are participating in a psychology experiment. You are at a computer in a room with the experimenter out of view and are told you will be interacting over a computer network with two other individuals, who are each in different rooms.

You are told that you and the other two participants— we will call them "***Person A***" and "***Person B***"— will each write a short essay expressing an opinion about a social issue of personal importance to each of you. For instance, you could write about marijuana legalization, gay marriage, abortion, if you are passionate about any of those issues. The specific topic is up to you. After everyone finishes their essays, everyone will then read each other person's essay and give that person feedback on the quality of his or her essay. Finally, everyone will read the feedback on their own essay and the feedback that the other participants received on their essays. Now, imagine that you and the other participants have completed these tasks, and are now examining the feedback. Specifically, imagine that you found ***Person A***'s and ***Person B***'s essays reasonable and gave both of them mildly positive feedback. ***Person A*** and ***Person B*** also, in general, gave mildly positive feedback (e.g., "This was a nice little essay. I could see why someone would think like this.").

However, ***Person A*** gave quite negative feedback on **[your]** ***Person B's*** essay: "I can't believe that an educated person would think like this. I sincerely hope that this person learns a thing or two."

Please indicate the extent to which you would feel **____ Person A** [**Person B**].
[The order of the emotions and which person considered first were randomized.]

angry, grateful, outraged, compassionate, happy, jealous, sympathetic, empathic, envious, mad

0=Not at all 1=Very little 2=Somewhat 3=Moderately 4=Very much 5=Extremely

Part 2:

Next, imagine that the experimenter asks you and the other participants to provide impressions of a certain recording of noise, as the experimenter may want to use that sound recording in future studies and needs to know how future participants will perceive it. The experimenter assigns you to administer the sound recording to ***Person A*** and ***Person B***, who will each rate how the sound made them feel. You are in charge of choosing how long each person hears the sound, and how loud it sounds for each person. Before you make your decision, you briefly hear the sound recording yourself: The sound is an irritating "white noise," similar to radio interference.

Please indicate how loud you would have ***Person A*** **(*Person B)*** listen to the sound recording:

1 (extremely quiet) – 10 (extremely loud)

Using the text box below, type in how many seconds you would have ***Person A*** **(*Person B*)** listen to the sound recording.