WHY WE HELP THE WRONGED: EMOTIONAL AND EVOLUTIONARY DETERMINANTS OF VICTIM COMPENSATION

Erik W. Thulin

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Cristina Bicchieri
Professor, Philosophy and Psychology

Graduate Group Chairperson

_____

Sara Jaffee
Professor, Psychology

Dissertation Committee
Barbara Mellers, Professor, Psychology and Marketing
Robert Kurzban, Professor, Psychology

ACKNOWLEDGMENT

Thank you to my parents, Judee and Chuck. Your steadfast advocacy and support, no matter my interests, allowed me to achieve.

Thank you to my partner, Azusa. You have been the bedrock of my life. It is on that foundation that this dissertation was built.

Thank you to my colleagues in the Behavioral Ethics Lab. You have served as an invaluable sounding board of incredibly diverse backgrounds, challenging my preconceptions.

Thank you to my committee, Barb and Rob. Your selfless offering of your extensive knowledge and experience has made me a better researcher.

Thank you to my advisor and mentor, Cristina. Your continued advice throughout my training has been invaluable. But that was the least of your contributions. You took me on as a student asking what you could do to realize my goals, opening doors I never knew were there. You shaped this dissertation, but also my life. For that, I will be forever grateful.

You are all my co-authors in this endeavor.

This material is based upon work supported by the National Science Foundation under Grant No. SES-1559320 and Graduate Research Fellowship Grant No. DGE-1321851.

ABSTRACT

WHY WE HELP THE WRONGED:

EMOTIONAL AND EVOLUTIONARY DETERMINANTS OF VICTIM COMPENSATION

Erik W. Thulin

Cristina Bicchieri

Why do third parties choose to help the victims of norm violations? In Chapter 1, we address this question at the emotional level. We show a relationship between environment and motivating emotion, in which moral outrage motivates the compensation of norm violation victims, whereas empathic concern drives compensation in other situations, at both the trait (Study 1) and state (Studies 2 and 3) levels. This finding presents a novel question for evolutionary psychology. Differing emotional drivers are taken to represent distinct underlying cognitive systems. While previous evolutionary models based on social insurance through indirect reciprocity can account for domain-general empathically driven compensation, they fail to address morally outraged compensation of norm violation victims. In Chapter 2, we extend two evolutionary models of punishment, showing how those same selection pressures may also account for victim compensation. We first propose the reputation-signaling hypothesis, under which compensators signal their community status and knowledge of local norms, making observers more likely to select them as future interaction partners. We also develop the norm stabilization hypothesis, in which compensators broadcast their endorsement of the violated norm, leading conditional conformists to continue to comply, thereby stabilizing the norm within the group. In Chapter 3, we develop and test empirical predictions of

both hypotheses. In Study 4, we find support for the joint prediction of both the reputation-signaling and norm stabilization hypotheses that compensation is increased when observed by others. In Study 5, we show that, consistent with the norm stabilization hypothesis, those who observe compensation of a victim of a norm violation are more likely to conform to that norm. In Study 6, we test the prediction of the reputation-signaling hypothesis that those who compensate are preferred as interaction partners to those who act similarly pro-socially, but not through compensation. Here we find mixed results, with compensators being preferred to those who show general pro-sociality, but less attractive than those who conform to an unrelated norm. Together, this work provides the first emotional and evolutionary account for the compensation of norm violation victims.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

INTRODUCTION

At 2AM On June 12, 2016 a man walked into Pulse nightclub in Orlando. Armed with a rifle and pistol, he went on a shooting spree at a crowd of more than 300 patrons. During the three-hour hostage standoff that followed, he killed 49 people and wounded 53 (The Washington Post, 2016). An unprecedented outpouring of support quickly followed. More than 7 million dollars were raised to compensate victims and their families from over 100,000 individual donors on the GoFundMe platform, the largest crowd-funded donation campaign in history (Rothaus, 2016).

Such compensatory behavior is not limited to crowd funding. In small-scale societies, hunter-gatherers with hunting windfalls have shown a willingness to compensate those with less successful hunts, and to be particularly generous to those who have a reputation of previously compensating (Gurven, 2004; Marshall, 1961). In modern large scale societies, the moral intuition that third parties ought to compensate victims has been incorporated into the mandate of the state. This belief is expressed in government sponsored health and unemployment benefits (Wendt, Frisina, & Rothgang, 2009). This intuition is also enshrined in many judicial systems, such as the New Zealand social insurance plan, which bans suing employers for injury. Instead, injury claims are paid from a central fund (Palmer, 1979). Similar legislation has been passed in the US, such as the September 11th Victim Compensation Fund, which collectively compensated terrorist attack victims for their loss (Harris, 2006). Although one might think of this as an extreme case, similar statutes have been enacted to address other harms, such as the National Childhood Vaccine Injury Act, the Price-Anderson Act to compensate those injured during nuclear

disasters, and the Black Lungs Benefits Act to compensate coal workers and their families (Mullenix & Stewart, 2002).

Third party intervention, including compensation, has drawn significant interest from across social science, with particular focus from psychology and behavioral economics. Experiments have shown third parties to be willing to compensate in a variety of experimental setups (Charness, Cobo-Reyes, & Jimenez, 2008; Chavez & Bicchieri, 2013; Leliveld, van Dijk, & van Beest, 2012). Across these studies, not only were third parties willing to compensate, they are willing to do so at a cost, paralleling the costs found in the natural environment (Baron, 2007).

These cases demonstrate that the drive to compensate exists across cultures, from hunter gathers to large-scale societies, permeating multiple levels of social interaction, from individual to individual exchanges to the legal regimes of nation states. Why do people choose to engage in compensation, even at a cost to themselves? It is this question which motivates this dissertation.

We approach the question of why people compensate from two interrelated levels of analysis: proximate emotional motivators and ultimate evolutionary selection pressures. Chapter 1 focuses on the emotional determinants. Psychologists have argued that the most proximate motivator of compensation is empathic concern for the victim (Coke, Batson, & McDavis, 1978; Batson, Duncan, Buckley, & Birch, 1981; Toi & Batson, 1982). Despite compensation occurring in many different social contexts, the literature has widely glossed over these differences, treating them all as the result of a single mental process. These contexts differ on at least one broad dimension: the cause of the victim's loss. While

some victims' losses can be attributed to bad luck, with no party to blame, others' losses are the result of a perpetrator's violation of the social rules governing that situation.

Punishment, another possible behavioral response to a norm violation, has also been attributed emotional motivations. However, whereas compensation has previously been accounted for as the result of empathic concern, punishment has been linked to experiencing moral outrage (Fehr & Fischbacher, 2004; Jordan, McAuliffe, & Rand, 2017; Nelissen & Zeelenberg, 2009). Despite their purportedly differing motivators, in the context of a norm violation, both compensation and punishment can serve similar cognitive goals, such as honoring the violated norm or giving people what they deserve (Carlsmith, Darley, & Robinson, 2002; Wenzel & Thielmann, 2006).

Up to this point, there has been no investigation into whether different emotional states may motivate compensation in different social contexts. Given the literature suggesting that other behavioral responses to norm violations (namely, punishment) are driven by moral outrage, and that both punishment and compensation can achieve similar cognitive goals in response to a norm violation, we propose that the broad characterization of compensation being driven by empathic concern across all domains may have been hasty. Instead, we tested a more nuanced account of the emotional motivators for compensation. We suggest that, while compensation may be driven by empathic concern when a loss is due to chance or a poor choice by the victim, moral outrage motivates people to compensate the victims of social norm violations, just as it motivates the punishment of perpetrators in the same context. Chapter 1 examines this proposal on both the trait (Study 1) and state levels (Studies 2 and 3).

The general pattern of results from Chapter 1, showing that moral outrage leads compensation in some contexts, whereas empathic concern leads to compensation in others, present an explanatory gap for evolutionary psychology. Previous work proposed the rationale for victim compensation as a form of efficient social insurance, supported through indirect reciprocity (Nettle, Panchanathan, Rai, & Fiske, 2011). However, if compensation is motivated by two different emotions in two different contexts, this suggests two different underlying mechanisms. Whereas empathic driven compensation across a wide variety of situations is quite consistent with the social insurance hypothesis, the finding that moral outrage drives the compensation of norm violation victims demands its own evolutionary rationale.

Chapter 2 takes on the challenge of providing an evolutionary account of the compensation of norm violation victims driven by moral outrage. Expanding on evolutionary models of punishment, we show how the same selection pressures which have been proposed to account for punishment of perpetrators may also account for the compensation of their victims. Specifically, we suggest that compensation may signal the compensator's quality as a future partner (reputation signaling hypothesis), or may help stabilize cooperative social norms within a group (norm stabilization hypothesis).

After developing these hypotheses, we then set about to test them. As the reputation signaling and norm stabilization hypotheses are the first evolutionary accounts of specifically the compensation of norm violation victims, we do not have an alternative model to make contrary predictions. Instead, we derive and test unique untested predictions of each of the accounts. As both of our proposed accounts argue that

compensation emerged for its informational value, they both predict that compensation should be sensitive to observation, which we test in Study 4. We then disentangle our two possible accounts, testing the norm stabilization hypothesis' prediction that people should be more willing to conform to a norm after witnessing compensation (Study 5) and the costly signaling hypothesis' prediction that participants should prefer to interact with compensators (Study 5) separately. Together, this work provides the first emotional and evolutionary account of the emotional and evolutionary psychology underlying the compensation of victims of norm violations.

CHAPTER 1

A long history of research in behavioral economics has demonstrated third parties' willingness to punish rule violators. This has been shown in a variety of games, including the prisoner's dilemma, ultimatum game, trust game, and dictator game (Charness et al., 2008; Chavez & Bicchieri, 2013; Fehr & Fischbacher, 2004; Kurzban, DeScioli, & O'Brien, 2007). This willingness has been found not only in the lab, but also in the field (Balafoutas, Nikiforakis, & Rockenbach, 2014; Mathew & Boyd, 2011). Although varying in size and prevalence, third-party punishment has been observed across a wide swath of cultures (Henrich, et al., 2010; Herrmann, Christian, & Gachter, 2008). A more recent line of inquiry has shown that third parties are also willing to *compensate* the victims of such rule violations in the ultimatum, dictator and trust games (Charness et al., 2008; Chavez & Bicchieri, 2013; Leliveld et al., 2012).

Across all these studies, not only were third parties willing to engage in punishment and compensation, but they were willing to *pay* to do so. This cost is crucial for their external validity, as both punishment and compensation are costly in the natural environment. When one engages in punishment, not only does one suffer the direct cost of necessary effort, but one is also exposed to the expected cost of retaliation, a risk born out in both the lab and field (Chagnon, 1988; Cinyabuguma, Page, & Putterman, 2006; Nikiforakis, 2008). Compensation bears the rather direct cost of losing whatever one chooses to compensate the victim with (Baron, 2007).

At the most proximate level, the willingness to engage in both these costly behaviors seems to be motivated by emotion. Much of past research on helping behavior has focused

on empathic concern, a constellation of emotions including feelings of sympathy, compassion, and tenderness as its primary driver (Batson et al., 1981). Empathic concern can be understood as an other-oriented emotional state, where one's own emotions are driven to be similar in valiance, although not necessarily identical, to those of someone in need (Batson, 1991). Empathic concern appears to lead to helping someone who received an unequal allocation in an economic game (Leliveld et al., 2012), volunteering to help a sick student (Coke et al., 1978), and even taking an electric shock to save a stranger from having to do so (Toi & Batson, 1982).

In contrast, anger, rather than empathic concern, is associated with a willingness to punish those who free-ride on the public good (Fehr & Gachter, 2001) or sanction someone who offers an unfair deal (Pillutla & Murnighan, 1996). People even become angry as third party observers of unfair treatment, leading them to engage in third-party punishment (Fehr & Fischbacher, 2004; Nelissen & Zeelenberg, 2009; Jordan et al., 2016).

The literature has treated compensation across various situations as generally similar psychologically. However, the contexts under which one might compensate vary dramatically. Broadly, one might compensate a victim when no one is at fault, or one could compensate victim of someone else's wrongdoing. Importantly, in the case of compensating the victim of someone's wrongdoing, compensation can serve some of the psychological functions previously identified as motivating punishment. Just as people punish to result in a more just outcome (Carlsmith et al., 2002), one can compensate a victim to give them what they deserve. Similarly, other work has suggested that people punish to restore the values of their community (Schroeder, Steel, & Woodell, 2003; Tyler

& Boeckmann, 1997; Wenzel & Thielmann, 2006) Just as punishment can reassert the social norm through costly signaling, the signal of compensating a victim of a violation can serve a similar purpose.

Given this symmetry in context and motivation between the punishment of norm violators and the compensation of their victims, we propose that there may be a similar symmetry in their emotional antecedents. We therefore suggest a finer grained understanding of the emotional motivators for compensation, involving *both* empathic concern and moral outrage. Specifically, we propose that the compensation of the victim of a norm violation is driven by the compensator's feeling of moral outrage, rather than their empathic concern for the victim. We therefore designed the following studies to test the hypothesis that moral outrage drives the compensation of victims, but only when the victim's loss was the result of a social norm violation.

In all three studies, we compare the effect of moral outrage and empathic concern on compensation across a variety of contexts, both through trait level correlations and experimental manipulation. The empathic concern hypothesis argues that empathic concern is the key emotional motivator for helping victims (Batson et al., 1981; Coke et al., 1978; Toi & Batson, 1982). However, based on moral outrage being the emotional antecedent to other behavioral responses to norm violations, such as punishment, we propose an alternative account, under which moral outrage, rather than empathic concern drives compensation. It is this divergence in accounts between the empathic concern hypothesis and our own that these studies aim to test.

In Study 1, we looked at the relationships between one's general dispositions to feel moral outrage and empathic concern (trait moral outrage and trait empathic concern), and willingness to compensate. Participants played a modified version of a hypothetical third party trust game (Charness, Cobo-Reyes, & Jimenez, 2008). They were assigned the role of a disinterested observer and shown the results of a modified trust game in which a player lost their endowment either due to the violation of a social norm, their investment going poorly, or chance. Participants were then given the opportunity to compensate this player's loss. In this study, we hypothesized that, contrary to the empathic concern hypothesis, trait level moral outrage would predict participant's willingness to compensate the victim of a social norm violation, but not when the loss was due to an investment gone awry or chance.

In Studies 2 and 3 we expanded our critique of empathic concern hypothesis to make a stronger *causal* claim for the role of moral outrage in driving compensation of norm violation victims. In Study 2, we experimentally manipulated participants' empathic concern and moral outrage while they took part in the modified trust games used in Study 1. In Study 3, we aimed to extend and replicate Study 2 using monetary incentives and a simplified compensation dependent measure. For these studies, we predicted that increasing moral outrage would lead to increased compensation of norm violation victims, but not those who experienced a loss for other reasons.

Study 1

In this study, we assessed the relationship between moral outrage and compensation at the trait level. We measured participants' willingness to compensate across a variety of

hypothetical contexts. In each context, another person lost money, either due to someone else's violation of a social norm, a bad investment, or chance. After observing this person losing money, the participant had the opportunity to compensate the victim for the loss by transferring some of his or her endowment to that person. We then measured each participant's general propensity to feel both moral outrage and empathic concern. We predicted that one's propensity to experience moral outrage would correlate with their willingness to compensate beyond their propensity of experience empathic concern, but only in the context of a norm violation.

Method

We recruited 241 participants (108 men, mean age of 33) from Amazon's Mechanical Turk (AMT) platform to participate in this study. We chose AMT to draw a more diverse sample than available from undergraduates. Previous work found that AMT samples are more diverse on age, geography, and ethnicity than undergraduate populations. In addition, the same work found responses from AMT to be at least as reliable as that gathered through traditional methods (Buhrmester, Kwang, & Gosling, 2011).

To measure the degree to which empathic concern and moral outrage influenced third party willingness to compensate across a variety of contexts, we used a series of modified hypothetical trust games with third party compensators. In the original trust game, the experimenter assigned participants to one of two roles, either that of the investor or the trustee. The investor received an initial endowment and could choose to transfer any of that amount to the trustee. The experimenter would then triple any amount transferred

by the investor. The trustee could then choose to send any portion of the tripled amount *back* to the investor (Berg, Dickhaut, & McCabe, 1995).

Using the original trust game as a foundation, we created three different interactions in which a participant may lose their endowment, due to either the violation of a reciprocity norm, a bad investment, or chance. These three situations served as three conditions in the study.

In the norm violation interaction, the experimenter endowed an investor with $10. The investor could then choose whether or not to transfer that $10 to the trustee. If the investor chose to keep the $10, the game ended. If they chose to transfer the $10 to the trustee, the experimenter quadrupled the amount to $40. At this point, the trustee could then choose to either keep the $40 or to return half ($20) to the investor. If the trustee chose to return half, the interaction ended. However, if the trustee chose to keep the entire $40, a third party observer, who was endowed with $10, was given an incentive compatible elicitation measuring the most the third party would be willing to pay to restore the investor to the original endowment of $10.

The bad investment situation was very similar to the norm violation situation, but with a single modification. Instead of the trustee having a *choice* of whether to transfer the $20 of the $40 to the investor, a randomizing device selected whether to return the $20. We chose the probabilities of an 80% chance of return of the $20 and a 20% chance of returning $0, which was known to all participants. These values were chosen to mimic the return rates in trust games of a similar setup (Fetchenhauer & Dunning, 2009). After observing the interaction, if the $20 was not returned to the investor, the third party

observer had the same choice as in the norm violation condition. Importantly, in this version of the interaction, if the investor chose to transfer their endowment, whether or not the $20 was returned to them no longer depended on the trustee conforming to a norm.

Finally, the chance interaction was similar to the bad investment interaction, but with one more modification. Instead of the investor having the *choice* of whether their $10 is transferred to the trustee (and then quadrupled by the experimenter), a randomizing device selected whether the $10 is transferred. We chose probabilities of a 50% chance of transferring the $10 and a 50% chance of not transferring the $10, which was common knowledge. These probabilities were again chosen to mimic the investment rates in trust games in a similar setup (Fetchenhauer & Dunning, 2009). As in the bad investment interaction, if the $10 was transferred to the trustee, a randomizing device then selected whether or not $20 is returned to the investor. If the $20 was not transferred back to the investor, the third party observer then had the same choice as in the previous two situations.

We randomly assigned each participant to one of the three interactions. After reading the complete description of one of the interactions, each respondent participated in a hypothetical instance of the interaction as the third party observer in which the investor's money was transferred to the trustee, but none was returned to the investor. After answering how much they would be willing to pay to restore the investor to their original $10, participants responded to inventories of trait propensity to feel moral outrage and trait propensity to feel empathic concern.

We adapted a four item trait moral outrage scale from previous work (Wakslak, Jost, Tyler, & Chen, 2007). For each item, participants expressed their agreement with a statement on a 7-point scale from "does not describe me well" to "describes me very well". Example statements included "I feel angry when I learn about people suffering from unfairness" and "I think it's shameful when injustice is allowed to occur".

We used the seven item Empathic Concern Subscale of the Interpersonal Reactivity Index to measure trait empathic concern (Davis, 1983). For each item, participants expressed how well it described them, on a five point scale of "does not describe me well" to "describes me very well". Items included "Sometimes, I don't feel very sorry for other people when they are having problems" and "I often have tender, concerned feelings for people less fortunate than me".

Results

188 participants (78% of the sample) correctly responded to at least nine of the ten comprehension questions asked throughout the instructions. In order to ensure high quality data, we used this subset in further analyses.

We found the four item trait moral outrage scale and seven item trait empathic concern scale to be highly internally reliable (α=.91 and α=.90, respectively). Additionally, trait level empathic concern and moral outrage were highly correlated with each other, $r(186)=.62$, $p<.001$. This high degree of correlation lead us to conduct all analyses of these variables controlling for the other in order to isolate the unique contribution of each.

For each condition, we analyzed the partial correlation between compensation and trait moral outrage controlling for trait empathic concern as well as trait empathic concern controlling for trait moral outrage. These results can be found in Table 1.

Table 1.

Partial correlations between compensation in each condition and empathic concern controlling for moral outrage and for moral outrage controlling for empathic concern

Condition	Empathic Concern	Moral Outrage
Norm Violation	-.152	.270*
Bad Investment	.397*	-.092
Chance	.012	-.063

Note. All values are partial Pearson correlation coefficients. *p<.05

Our key prediction was that moral outrage would predict compensation in the Norm Violation condition, while not doing so in the Chance and Bad Investment conditions. We see this supported in the Moral Outrage column of Table 1, where, controlling for empathic concern, moral outrage was significantly correlated with compensation in the Norm Violation condition, $r(65)=.27$, $p=.027$. Also importantly, we see that, controlling for empathic concern, moral outrage predicted compensation in neither the Chance condition $r(56)=-.063$, $p=.636$ nor the Bad Investment condition $r(58)=-.091$, $p=.49$. In fact, both of these non-significant effects had a negative sign.

We observed that empathic concern, controlling for moral outrage, was correlated with compensation in the Bad Investment condition $r(58)=.40$, $p=.002$. However, empathic concern was not correlated with compensation in either the Norm Violation condition $r(65)=-.15$, $p=.22$ or the Chance condition $r(56)=.01$, $p=.93$.

Discussion

Past research includes numerous examples of helping behavior correlating with empathic concern, across a variety of contexts, from volunteering to help a sick student to paying to compensate someone who received an unfair allocation in a behavioral game (Coke et al., 1978; Leliveld et al., 2012; Toi & Batson, 1982). Our initial finding that the dispositions to feel moral outrage and the disposition to feel empathic concern are highly correlated suggests an important caveat when interpreting earlier studies: as these experiments did not address moral outrage as a covariate, it is possible that effects interpreted as being driven by empathic concern may in fact have been driven by an important third variable, namely moral outrage. Study 1 investigated the plausibility of this claim, looking at the unique contributions of trait empathic concern and trait moral outrage across three contexts. In support of this past literature, we find that empathic concern *does* maintain a unique correlation with compensation controlling for moral outrage, but only in particular contexts, namely in the Bad Investment condition where someone makes a risky decision and suffers a loss. We do not see any unique correlation between compensation and empathic concern in the Chance condition, where all transfers were randomized. Although not directly linked to the questions at hand, future work may illuminate what differences between the Bad Investment and Chance conditions lead to the differing effect

of empathic concern, and perhaps answer what motivations may be present in compensating the victims in a chance-like scenario.

Our focal question for this study asked whether a propensity to feel moral outrage was related to a willingness to compensate, and whether that effect was limited to the case of social norm violations. The analysis of the correlations of moral outrage with compensation, controlling for empathic concern, across the various conditions suggest the answer to both questions is yes. In the case of the social norm violation, we find that compensation correlated with moral outrage, controlling for empathic concern. In addition, we find that moral outrage did not correlate with compensation in the other two conditions. This provides evidence that the empathic concern hypothesis, that helping in general is due to empathic concern is not sufficient: a more fine grained account is required. However, this study only looked at trait emotional dispositions, and therefore assessed correlations. In order to better understand the causal effect of emotion on compensation, direct manipulation is required.

Study 2.a

Study 2 extended the findings of Study 1 from the trait domain into that of emotional states. Study 2.a investigated the relationship between participants' current level of moral outrage and the degree to which they were willing to compensate. Whereas the previous study relied on correlational relationships with trait variables, we were able to manipulate emotional states, allowing for stronger causal claims. In this study, we manipulated the amount of moral outrage a participant experiences using video inductions. We then assessed their willingness to compensate across the three hypothetical situations used in

Study 1. Finally, we measured the degree to which each participant was currently experiencing moral outrage and empathic concern. We predicted that those led to experience moral outrage would be willing to compensate more than those who were not, but that this effect would be limited to the norm violation context. Additionally, we predicted that, controlling for empathic concern as a covariate, experienced moral outrage would mediate the effect of the video induction on willingness to compensate.

Method

We recruited 990 participants (471 men, mean age of 33) from the AMT platform to participate in this study.

We experimentally manipulated moral outrage, measuring its effect on compensation across the three hypothetical situations developed in Study 1: norm violation, bad investment, or chance. Each participant read instructions describing the interaction, while answering a series of comprehension questions throughout.

After reading the instructions, but before being told what role in the interaction they would be assigned to, participants watched a short video, serving as the manipulation of moral outrage. This manipulation took advantage of people's tendency to attribute arousal states such as anger to whatever stimulus they are currently being exposed to (Schachter & Singer, 1962). Those assigned to moral outrage watched a short video of a boy being attacked by a bully, which past work identified as significantly increasing moral outrage (Lerner, Goldberg, & Tetlock, 1998). Participants assigned to low moral outrage watched a video of abstract line patterns, previously found to be emotionally neutral (Gross & Levenson, 1995).

After watching one of the two videos, all participants were assigned to the role of the third party observer and asked how they would respond if the investor's funds were transferred to the trustee, but none were returned to the investor. Participants were then given the same hypothetical version of an incentive compatible elicitation used in Study 1, measuring their willingness to pay to restore the investor to their original $10. After giving their responses, participants then answered a series of questions measuring their current levels of empathic concern and moral outrage.

The four item state moral outrage scale was used in previous work for the same purpose (Piazza, Russell, & Sousa, 2013). For each item, participants rated the degree to which they agreed on a five point scale from "Strongly Agree" to "Strongly Disagree". Example items included "I feel angry" and "I feel outraged". We adapted three items from the Empathic Concern Subscale of the Interpersonal Reactivity Index used in Study 1 in order to measure state empathic concern. For each item, participants rated how much they agreed on the same five point scale used for the state moral outrage items. Items included "I feel sorry for Person A" and "I was disturbed by what happened to Person A", Person A being the investor in their interaction.

Results

754 (76%) of participants correctly responded to 9 of the 10 comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

Both the four item state moral outrage scale and the three item empathic concern scale showed high degrees of internal reliability (α=.958 and α=.847, respectively). Using the moral outrage scale as a manipulation check, we found that moral outrage was

significantly manipulated in the norm violation situation t(230)= 5.18, p<.001, chance situation t(262)=4.51, p<.001, and the bad investment situation t(256)=3.27, p=.001. These effects ranged in size across situations from *d*=.41 to *d*=.68, demonstrating a medium sized effect of the video manipulation on moral outrage.

We report mean levels of compensation across conditions in Figure 1. In the norm violation situation, we found that those who watched the moral outrage video were willing to pay significantly more to compensate (M=4.09) than those who watched the neutral control video (M=3.16), t(230)=2.41, p=.017. We then tested whether a subject's feeling of moral outrage mediated this effect, the results of which can be found in Figure 2. In this and all following mediation analyses, we ran non-parametric bias-corrected boostrap analysis (Hayes & Preacher, 2014) with 10,000 resamples. Controlling for empathic concern as a covariate, moral outrage significantly mediated the effect of the video manipulation on the amount participants were willing to pay to compensate *B*=.31, 95% CI=.08 to .65.

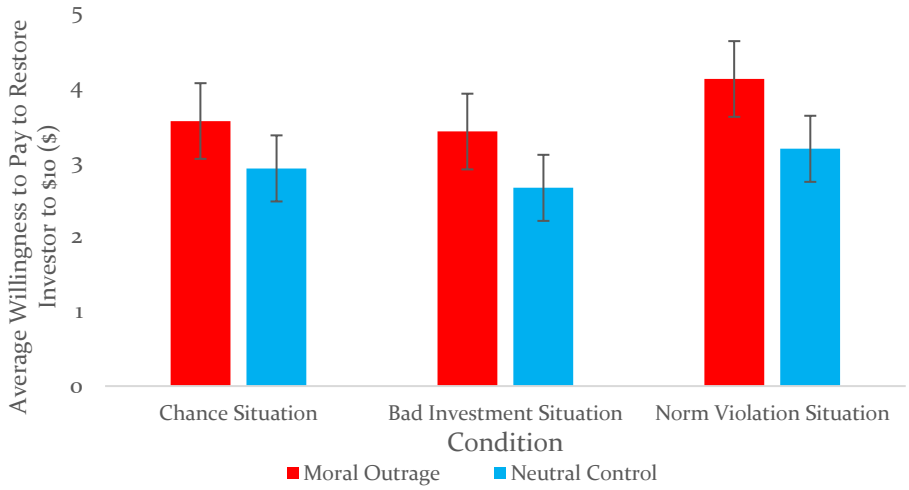

Figure 1: Average willingness to pay of participants to restore investor to $10 by condition, error bars as 95% confidence intervals

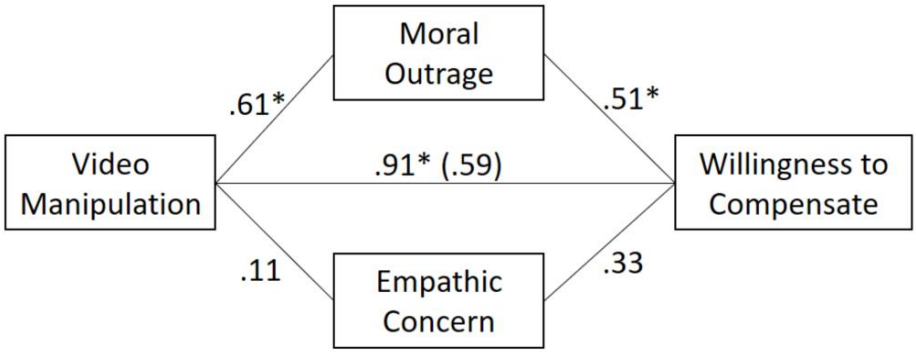

Figure 2. Mediation model for the relationship between the video manipulation and willingness to compensate in the norm violation situation, as mediated by moral outrage and empathic concern. Relationship between the video manipulation and willingness to compensate, controlling for empathic concern and moral outrage, is in parenthesis. *p<.05

Participants who watched the moral outrage-inducing video in the bad investment situation also compensated significantly more (M=3.53) than those who watched the neutral control video (M=2.64), t(262)=.42, p=.042. We observed in that situation that empathic concern also differed significantly between the moral outrage and control video conditions t(262)=2.40, p=.017. Mediation analysis showed that while moral outrage was *not* a significant mediator of the effect of the video on compensation, *B*=.01, 95% CI=-.22 to .21, empathic concern *was* a significant mediator, *B*=.19, 95% CI=.02 to .38.

In the chance situation, those who watched the moral outrage video also compensated significantly more (M=4.09) than those who did not (M=3.16), t(256)=2.39, p=.017. However, similar to the bad investment situation, controlling for empathic concern as a covariate, moral outrage was *not* a significant mediator of the effect of the video on compensation, *B*=-.17, 95% CI=-.25 to .07.

Discussion

The finding that increased moral outrage led to increased willingness to compensate in the norm violation situation provides support for the causal role of moral outrage in compensating the victims of social norm violations. The finding that, controlling for empathic concern, moral outrage mediated the effect of the video manipulation on compensation further bolsters the claim of moral outrage's causal role.

We did not predict that compensation would be higher in the bad investment and chance situations after watching the moral outrage inducing video, which led us to conduct further tests to better understand those results. We observed that, although we chose the video due to its limited effect on other emotions, it also significantly affected empathic concern in the bad investment situation, which allowed for the possibility that it was the change in empathic concern, rather than moral outrage, which drove the effect. To test for this, we used mediation analysis, allowing for both empathic concern and moral outrage to serve as mediators of the video's effect on compensation in the bad investment situation. The finding that, in the bad investment situation, empathic concern, and not moral outrage, mediated the effect of the video on compensation is consistent with empathic concern, rather than moral outrage, driving compensation.

Similarly, we ran a mediation analysis in the chance situation, testing the degree to which moral outrage mediated the effect of the video on compensation. Similar to the bad investment situation, we did not find moral outrage to be a significant mediator. Here we see a parallel of Study 1, where we found support for moral outrage *not* being a

determining factor of compensation in the chance situation, but these data do not speak to what may actually be the emotional determinants.

One important concern to address is that of a demand effect. Demand effects can occur when participants are aware of what the experimenter expects them to do, and choose to conform to that expectation (Orne, 1962). Although the effect of deception, both within a given study and on the public good of participant pool perceptions, is a hotly contested and ongoing debate (Cook & Yamagishi, 2008; Hertwig & Ortmann, 2008), we chose to adopt the proscription of deception throughout this studies. As a result, when participant emotional state was manipulated, no misleading cover story was given to ensure that the intent of the manipulation was obfuscated. This raises the question of whether the observed effects could be due to demand.

In order to assess the plausibility of a demand effect, we need to first establish what conditions are necessary for it to have occurred, and what patterns of data we would predict under this alternative account. As this study was conducted between, rather than within subjects, participants would first need to infer that the difference between conditions was what video was played. Participants would then need to correctly guess that the experimenter prediction was for those having seen the bully video, rather than the control video, to be more willing to compensate. After determining what element differed between conditions as well as the predicted direction of effect, participants would need to choose to conform to that deduced expectation.

We then need to consult the pattern of results and assess the degree to which they are consistent with the demand effect rationale. We found that moral outrage mediated the

effect of the video on compensation in the Norm Violation condition, but in neither the Bad Investment nor the Chance conditions. For this pattern to occur under the demand effect account, we must not only accept the assumptions listed above, but also that participants in the Norm Violation condition who reported moral outrage would figure that they should compensate more, but that participants in the Bad Invest and Chance conditions who experienced more moral outrage would figure out that, given the parameters of the game they were assigned to, the experimenter would *not* predict that the video's effect should be mediated by a feeling of moral outrage, and would choose to compensate less frequently.

These assumptions are possible to hold. However, we find the alternative hypothesis that empathic concern and moral outrage have different effects on compensation in different contexts to be more plausible. Consistent with this conclusion, the Gross and Levenson (1995) video manipulations have previously been used to manipulate emotional state without a guise (Gross & Levenson, 1997; Drouvelis & Grosskopf, 2016; Fredrickson & Levenson, 1998).

Study 2.b

Study 2.b closely mirrors the design of study 2.a, but focuses on the role of state empathic concern rather than moral outrage. In this study, empathic concern towards the person who lost their money was manipulated by having the participant either write a response to a prompt asking them to take the perspective of the person who lost their money, or to neutrally describe the interaction. Each respondent then participated in one of the three situations described in Study 1. We predicted that, consistent with previous work, those

who responded to the high empathic concern prompt would compensate more in the situations not involving a norm violation, but that this pattern would not be present in the norm violation situation. Additionally, we predicted that, controlling for moral outrage, empathic concern would mediate the effect of the perspective taking manipulation on compensation in the non-norm violation situations.

Method

We recruited 998 participants (472 men, mean age of 34) from the AMT platform to participate in this study.

The design of Study 2.b closely mirrored that of 2.a, with the key difference being our manipulation of empathic concern rather than moral outrage. Whereas anger is experienced as a general emotional state, empathic concern is, by its very nature, expressing concern *for* a particular person, which did not allow us to use a video manipulation. Instead, after reading the rules to the interaction, being assigned to their role as the third party, and seeing that the investor did not receive any money back, we had participants write in response to one of two prompts. In the control conditions, we asked participants to "objectively describe what has happened in the interaction so far". In the empathic concern conditions, we asked participants to "describe the feelings and emotions Person A may be feeling right now".

Results

769 (77%) of participants correctly responded to 9 of the 10 comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

Using the empathic concern scale as a manipulation check, we found that the prospective taking prompt lead to higher empathic concern relative to the control in the norm violation situation, t(269)=2.92, p=.004, the investment situation, t(244)=3.34, p=.001, and the chance situation, t(250)=3.40, p=.001. These effects were moderate in size (*d*=.36 to *d*=.43).

We report mean levels of compensation in Figure 3. Those in the bad investment situation who received the empathic concern prompt (M=4.03) compensated significantly more than those who received the objective prompt (M=3.05). In the bad investment situation, controlling for moral outrage as a covariate, participants' level of empathic concern significantly mediated the effect of the prompt manipulation on compensation, D=.45, 95% CI= .1634 to .7010.

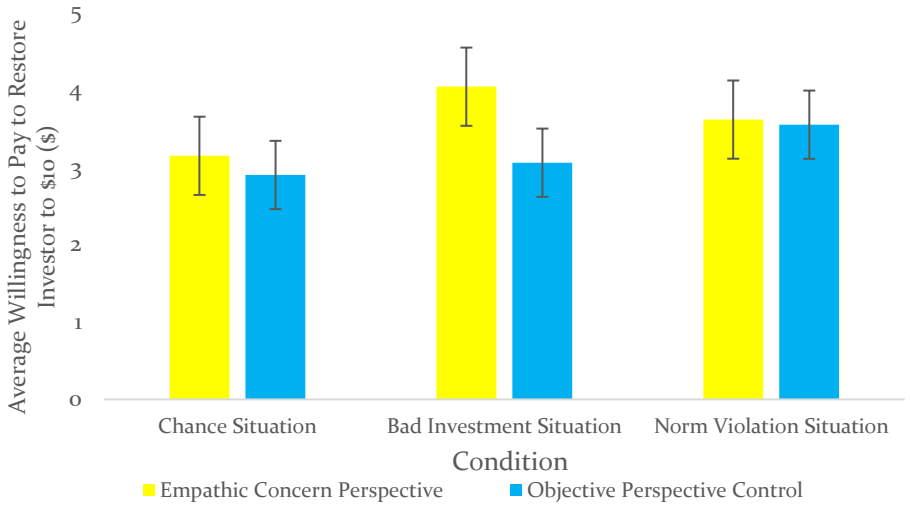

Figure 3: Average willingness to pay of participants to restore investor to $10 by condition, error bars as 95% confidence intervals

In the chance situation, we did not find that those who responded to the empathic concern prompt (M=3.13) compensated significantly more than those who responded to

the objective prompt (M=2.89), t(250)=.822, p=.41. Similarly, in the norm violation situation, we found no significant difference in compensation between those who received the empathic concern prompt (M=3.61) and those who received the objective prompt (M=3.54), t(269)=.132, p=.90.

Discussion

The empathic concern prompt leading to higher compensation in the bad investment situation supports the hypothesis and results from previous studies that empathic concern can drive compensation behavior. Mediation analysis further buttresses this finding, showing that empathic concern mediates the effect of the written prompt on compensation.

Consistent with our findings in Study 1, we did not find a significant effect of empathic concern on compensation in the chance situation. This provides additional motivation for further investigation into what may be driving compensation in this context. We also do not find a significant effect of empathic concern on compensation in the norm violation situation, consistent with our general hypothesis that moral outrage, rather than empathic concern, drives compensation in the context of norm violations.

Participants in Study 2.b were subject to similar possible demand characteristics as those in Study 2.a. We again suggest that our proposed account is more plausible than subjects inferring the particular manipulation, our particular expected direction of effect, and choosing to comply with that effect. The particular pattern of results in Study 2.b do also not lend themselves to a demand explanation. Note that the effect of the empathic concern video manipulation is found in the bad investment situation, but not the norm

violation situation. For this to be the case under the demand account, not only would subjects in the bad investment situation need to deduce that those who were told to write about the investor's emotions were expected to compensate more than those who were told to describe the situation, but those in the norm violation condition would need to deduce that participants in their situation were expected to *not* be effected by the manipulation.

Study 3

We designed Study 3 to replicate and generalize the finding of Studies 1 and 2.a that, in the case of a social norm violation, moral outrage correlated with (Study 1) and drove (Study 2.a) participants' willingness to compensate. This study had two manipulations. First, participants were assigned to either the norm violation or bad investment situations previously described in Study 1. Second, participants were assigned to either a moral outrage or neutral emotional video manipulation described in Study 2.a. We made two other key modifications from Study 2.a. First, participants interacted with each other for actual money rather than responding to hypothetical situations. Second, participant feedback suggested that the willingness to compensate measure used in Studies 1 and 2 was complex and therefore difficult to understand. We therefore substituted a simple transfer with multiplier as the dependent measure to improve participant comprehension. We predicted that those who watched the moral outrage inducing video would compensate more than those who did not, but only in the social norm violation situation.

Method

We recruited 502 participants (243 men, mean age of 38) from the AMT platform to participate in this study.

Participants were divided into two phases. Those in Phase 1 read a description of a trust game, similar to those used in the previous studies. In the norm violation situation, investors were endowed with $0.50 and trustees with $0.00. The investor could choose to either keep their $0.50 or transfer it to the trustee. If transferred, the experimenter tripled the amount to $1.50. The trustee then had the option of whether to keep the entire $1.50 or to return $0.75 to the investor. As in the previous studies, the bad investment situation mirrors the norm violation situation, aside from one variation. Instead of the trustee *choosing* whether or not half the transfer was returned, a randomizing device selected, returning half the endowment 80% of the time and none of the endowment 20% of the time. So as not to deceive participants, they were informed that the choices of future participants may impact their payoffs. Phase 1 was run until, for both the norm violation and bad investment situations, an investor chose to transfer their endowment to the trustee and the trustee chose not to return the sum. These final pairs were used as the focal dyads.

After establishing the focal dyads, all further participants were assigned to Phase 2. Each participant in Phase 2 read a description of the trust game outlined above. Participants were told that they were assigned to the role of a third party for an investor and trustee pair and given an endowment of $0.75. They were told that if the investor chose to transfer their $0.50 to the trustee but $0.75 was not returned from the trustee to the investor, they

would have the opportunity to transfer any amount of their $0.75 to the investor, and that the amount they chose to transfer would be doubled by the experimenter.

After reading these interaction instructions, Phase 2 participants were shown one of the two videos used in Study 2.a, to either induce moral outrage or serve as a neutral control. After watching the video induction, participants were shown the result of one of the focal dyads, in which the investor chose to transfer to the trustee and either a randomizing device or the trustee selected not to return half the endowment, depending on condition[1]. After seeing the result, participants then chose how much of their endowment to transfer to the investor, which was then doubled by the experimenter.

After making their selections, participants responded to the state moral outrage and empathic concern scales used in Studies 2.a and 2.b. Participants were immediately paid their $0.50 show up fee, and then paid their bonus amounts five to seven days later.

Results

385 (77%) of participants recruited for Phase 2 correctly responded to 4 of the 5 comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

[1] The method of using a focal dyad for all future decision has previously been used previously to maintain non-deception, as nothing false is told to participants, while increasing the efficiency of the study by minimizing the number of subjects necessary to achieve adequate power (Kurzban et al., 2007).

Using the four item state moral outrage scale as a manipulation check, we observed that the moral anger video significantly increased the level of moral outrage relative to the control video, $t(486)=4.87$, $p<.001$. This is a moderately sized effect ($d=.44$).

The mean compensation values across conditions are shown in Figure 4. We predicted that, in the norm violation situation, those participants who watched the moral outrage-inducing video would compensate to a greater amount. However, the difference observed was small and non-significant, $t(196)=.40$, $p=.687$.

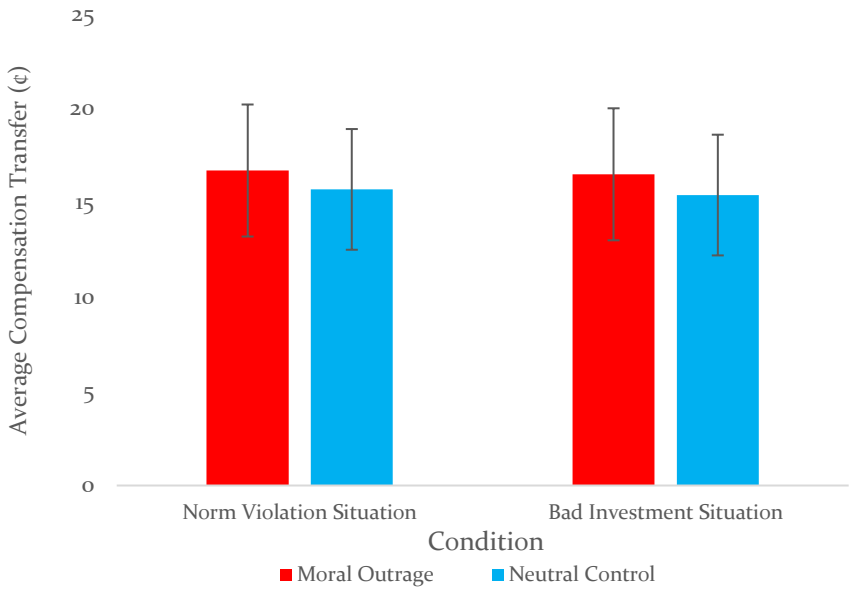

Figure 4: Average amount transferred from participants to investor by condition, error bars as 95% confidence intervals

Due to this surprising result, we also investigated the partial correlations between moral outrage and compensation, controlling for empathic concern, in both the norm violation and bad investment contexts. Controlling for empathic concern, we found that moral outrage was significantly correlated with compensation in the norm violation context, $r(197)=.15$, $p=.03$. This differed from the bad investment context, in which we did not find a

significant relationship between moral outrage and compensation, controlling for empathic concern, $r(184)=.09$, $p=.24$.

Discussion

The lack of an effect of the video manipulation on compensation was surprising, and inconsistent with the results of Studies 1 and 2.a. There were two key differences between the previous studies and Study 3, which may have affected the result. The first, and most concerning, is that the effect may exist for hypothetical exchanges but does not generalize to exchanges involving actual incentives. There is reason to be suspicious of this possibility, as past research has shown that subjects drawn from AMT respond similarly to hypothetical games as they do to those involving actual money, including in the specific context of trust games (Amir, Rand, & Gal, 2012).

A second difference between the previous studies and Study 3 was the elicitation and measurement of compensation. In the previous studies, we used a hypothetical incentive compatible elicitation of the most one was willing to pay to restore the investor to their original endowment. We gave each participant a series of binary choices, asking if they would be willing to pay X in order to restore the investor to $10, where X ranged from $1 to the third party's entire endowment of $10. After making their choices, one of the ten choices was randomly selected and carried out (for example, if the "Would you be willing to pay $3 of your $10 to make person A end with $10?" question was selected and the participant chose "Yes", then the participant would have $3 deducted from their endowment, and the investor would receive $10).

We chose this original method because willingness to pay has a high degree of granularity as compared to a single choice (for example, only asking would you pay $2.50 to restore the investor to $10). It also measures the implicit lowest compensation tradeoff ratio that a participant sees as making the transfer worthwhile, which we find to be a compelling proxy for one's willingness to compensate. For example, being willing to transfer $4 but not $5 to restore the investor to $10 implies the minimum acceptable compensation tradeoff ratio between 2.5 and 2. This is in contrast to choosing an amount to transfer with a fixed multiplier, which the interpretation of is much more ambiguous. As opposed to the willingness to pay measure, one cannot impute the minimum acceptable multiplier (as the multiplier is held constant). Instead, the amount transferred could indicate what the third party thinks would be the *correct* amount of compensation, rather than to what degree the third party cares whether the investor receives that correct compensation.

Although the willingness to compensate measure had these desirable properties, feedback from participants in Studies 1 and 2 suggested that the method was very difficult to understand, and at a minimum, cognitively taxing. As we were particularly interested in the emotional determinants of the compensation decision, we chose to simplify the compensation measure in Study 3 by simply asking how much of their endowments participants wished to transfer to the investor, with a 2x multiplier. The lack of an effect in this case may therefore be because, while previous dependent measures assessed to what degree the third party wanted to restore the investor to their original state, the current measure may be assessing what amount the third party thinks is the correct amount of compensation, which may be less subject to influence from moral outrage.

These concerns are partially assuaged by correlations within the data being consistent with Studies 1 and 2.a. Namely, the finding that moral outrage correlated with compensation in the norm violation situation but not in the bad investment situation, controlling for empathic concern, is the same pattern observed in Studies 1 and 2.a. This is consistent with the general hypothesis that moral outrage makes a significant unique contribution to the compensation of norm violation victims.

General Discussion

Taken together, these studies begin to reveal a richer landscape of emotional determinants of victim compensation than was previously identified. Studies 1 and 2.a found that on both the trait and state levels, moral outrage was associated with a willingness to compensate victims of social norm violations beyond the effect of empathic concern. In fact, when we controlled for moral outrage, or directly manipulated empathic concern, the data revealed no significant effect of empathic concern on the compensation of victims of social norm violations.

Also as predicted, the effect of moral outrage on compensation appears to be domain specific. We found no significant relationship between a propensity to feel moral outrage and willingness to compensate when a loss was due to chance or a bad investment in Study 1. Despite finding significant differences in willingness to compensate in both the chance and bad investment situations in Study 2.a, we found that moral outrage mediated neither of these effects. This result was consistent with Study 1, suggesting that moral outrage was not involved in driving compensation in these contexts.

Our finding in Study 3 that increasing moral outrage did not increase compensation in the norm violation context contrast with the pattern of results in Studies 1 and 2.a. One possible explanation for this discrepancy was the change in dependent measure. Whereas Studies 1 and 2 measured the most one is willing to pay to restore the investor to $10 (effectively measuring the lowest compensation trade-off ratio the third party is willing to accept), Study 3 measured the amount the trustee chose to transfer. The latter is at least partially determined by the amount a participant feels is the *correct* amount to transfer rather than the degree to which they want the recipient to get that amount. It is possible that, while the degree to which one wants a recipient to get the correct amount *is* influenced by moral outrage, the correct amount itself is not. Future work may assess this by manipulating moral outrage and assessing its effect on willingness to pay to compensate as compared to the amount one is willing to compensate. Despite this inconsistent finding, even in this study we found that moral outrage, controlling for empathic concern, correlated with compensation in the norm violation situation but not the bad investment situation, consistent with the previous pattern of results.

An unexpected but interesting result emerged when evaluating the relationship between empathic concern and compensation in the chance situation. In the bad investment situation in both Studies 1 and 2.b, we found relationships between empathic concern and willingness to compensate. However, in the chance situation, we found no such relationships between compensation and empathic concern. Further work is required to understand the distinguishing features between these two cases, and what other emotional determinants may be driving compensation when losses are due to chance.

At first glance, this general pattern of findings seems inconsistent with previous work demonstrating a relationship between empathic concern and third party compensation of those who receive low offers in a dictator game (Leliveld et al., 2012). However, there are two possible ways to reconcile these findings. First, as reported in Study 1, there is a high correlation between moral outrage and empathic concern, which points to the importance of controlling for one to understand the influence of the other. As this previous work did not include such controls, it is possible that moral outrage, as a latent third variable, may account for the results. Second, other work has shown that people do not have strong personal beliefs of what divisions one *should* make in the dictator game, which is critical for the existence of a social norm (Bicchieri, 2006). As no norm may exist in the dictator game situation, and therefore none may be violated, it would be reasonable for empathic concern, rather than moral outrage, to motivate third parties to compensate.

This work was confined to artificial contexts using behavioral games. However, if the results can be shown to generalize more broadly, it may help us better understand the motivations for charitable giving, and therefore have implications for those soliciting such donations. Previous studies have suggested empathic concern drives people both to volunteer (Davis, et al., 1999) as well as engage in charitable giving (Bekkers, 2006). Our results suggest that picture may be incomplete, and that the degree to which people are motivated to help may be driven by different emotions in different contexts. Future work may assess the degree to which moral outrage may be a motivating factor for volunteerism and charitable giving when the target is the victim of a norm violation, and assess whether messaging focusing on this theme in those contexts is effective.

Finally, this work has focused on the proximate emotional determinants of victim compensation. Researchers in evolutionary psychology have often seen emotions as proxies for underlying cognitive systems (Fessler & Haley, 2003; Haidt, 2003). When two separate emotions are shown to drive behavior in two different contexts, this suggests different evolved mechanisms in play. By revealing a novel emotional motivator for the prosocial compensation of norm violation victims, we are left with the question of what mechanism drives this behavior, and what can we deduce about its ultimate origins. Chapter 2 develops possible answers to these questions, and poses novel hypotheses by which to test them. Chapter 3 carries out studies to test these hypotheses in order to better understand the underlying mechanisms motivating the compensation of norm violation victims.

CHAPTER 2

The results from the studies in Chapter 1 provide a more fine-grained understanding of the emotional motivations for compensatory behavior. Consistent with many previous studies, we found that empathic concern for the victim drove compensation in some contexts. However, this was not as universal as previously reported. We found that when someone experienced a loss due to the violation of a social norm, it was moral outrage, rather than empathic concern, which drove third parties to compensate their loss.

Evolutionary theorists have addressed the question of why one might compensate someone else's loss. They have paid particular attention to the implication of diminishing marginal fitness benefits of resources for the incentive to help (Nettle et al., 2011). The key insight of these models is that, as one acquires more of a particular resource, the marginal benefit of that resource tends to decrease. For example, the first calorie of perishable elk meat provides significantly greater fitness benefit to the hunter than the 100,000th. This creates a unique opportunity for cooperation over time. When the amount of resources one acquires (or loses) varies over time, individuals can improve their overall fitness outcomes by offsetting their losses with the excess from their gain periods. However, as many resources are perishable or hard to defend in large quantities, it is costly if not impossible to engage in this intertemporal offsetting within the individual. This creates a cooperative context in which individuals would benefit if they were able to receive help when they have relatively little, at the cost of giving help when they have relative excess.

Although this dynamic is to some degree applicable to all situations involving generous behavior, it is particularly important in the context of the compensation of momentary

victims. Taking the example of disease, anyone in a community could fall ill. Such a situation creates a great momentary difference in the benefit of aid. While a fully healthy adult would benefit little from what trivial help the ill person could provide at that moment, the ill person may benefit dramatically by the aid of others. After one heals, she may have the opportunity to offer such aid to another. It is precisely this boom-and-bust in resources and need that create the fertile ground for the emergence of a cooperative tendency to compensate victim's losses, conditional on the mutual expectation that, if you do so, you will be helped in a similar situation in the future.

While this dynamic demonstrates the unique cooperative opportunity presented in victim compensation, it does not suggest a particular evolutionary solution. Theorists have pointed to reciprocity, and indirect reciprocity in particular, as perhaps the force driving the evolution of compensation (Nettle et al., 2011). They suggest that those who compensate build up reputations as compensators, and when one experiences a loss, others compensate that loss conditional on whether the present victim was a previous compensator. Indeed, patterns outside the lab suggest that may be the case: in small scale societies, those who are sick but have a good reputation for sharing in the past receive particularly generous shares of food to compensate for their illness (Gurven, 2004).

The efficient helping hypothesis described above treats compensation as if it were the result of a single evolved system, paying no attention to varying motivators across contexts. However, the results of Chapter 1 suggest that compensation is driven by two different emotions in two distinct contexts. Dating back to Darwin's *The Expression of Emotions in Man and Animals* (1872), evolutionary theorists have taken emotions as

indicators of specific evolved mechanisms (Tooby & Cosmides, 2008; Fessler & Haley, 2003; Haidt, 2003). Under this framework, if compensation is driven by empathic concern when losses are not intentionally inflicted, but moral outrage when losses are due to norm violations, these must be driven by distinctly evolved systems. Of these two systems, the efficient helping hypothesis best fits empathic concern-driven helping, as it is elicited broadly by seeing one in need, rather than moral outrage driven compensation, which would only be found when a norm the compensator cares about is violated (Haidt, 2003). This therefore leaves a hole in the literature. If our best understanding of third party compensation can explain empathic concern driven compensation, but cannot account for moral outrage driven compensation, what led to its evolution?

As both social norm violation victim compensation and perpetrator punishment appear to be driven by the same proximate emotion, namely moral outrage, we suggest that they are both the outcome of a single underlying system. We therefore propose a social norm violation response system, under which one emotion, moral outrage, motivates two linked behaviors, violator punishment and victim compensation.

This chapter aims to provide an evolutionary account of this norm violation response system generally, and norm violation victim compensation specifically. We do so in two parts. In the first section, we review the existing literature on evolutionary models of third-party punishment, with a particular focus on those which may be also be applicable to victim compensation, including reputation and cultural group selection models. In the second section, we expand these models to provide a novel evolutionary account of third party compensation, to then be tested in Chapter 3.

The Evolution of Third Party Punishment of Norm Violators

Cooperation among non-kin is a defining feature of human social life. However, cooperation presents an evolutionary conundrum: if we understand evolution to favor those who do what is in the interest of their inclusive fitness, how might it select for the cooperative tendencies that appear so prevalent in human interaction? Many theorists found punishment to be a suitable candidate for sustaining cooperation. Assuming the probability of punishment multiplied by the cost to the perpetrator exceeds the benefit derived from failing to cooperate, punishment does indeed incentivize cooperation. However, punishment is less a solution to cooperation as it is another evolutionary quandary. Punishment often has associated costs, whether they be the direct cost of punishing, or the indirect costs of reprisal. This creates what is known as a second order free-rider problem: punishment conveys the group level benefit of driving others to cooperate, but this comes at the expense of the punisher, making punishment itself a cooperative act in need of explanation.

A wide variety of models for the emergence of third-party punishment have been proposed. However, all of these models are broadly based on either reputation or on multilevel selection (Nowak, 2006). In this section, we review these two classes of models, with a particular focus on those which will be built upon in the next section to propose an evolutionary account of morally outraged compensation.

Reputation-Based Models of Punishment

Our understanding of the evolutionary benefits of reputational effects rests on our understanding of costly signaling. Zahavi (1975) first used costly signaling to explain the

evolution of the colorful plumage of the male peacock. As these large colorful tails are biologically costly, one must ask how they could be selected for. Zahavi proposed that the tails were selected for, precisely because only those who possessed hidden quality were able to take on the cost of producing them. This created a correlation between plumage and underlying quality. This correlation functioned as an honest signal, under which those females who chose the males with more plumage gained mates of higher intrinsic quality, thereby producing offspring of higher quality. This logic was concurrently formalized in economics (Spence, 1973), showing that education could function as a signal of employee quality, assuming that the opportunity cost was lower for those of high employee quality than those of low. This model was generalized and returned to biology via evolutionary game theory, showing that costly signaling was evolutionarily stable if those of high intrinsic quality paid a lower cost to signal (or experienced a greater benefit) than those of lower quality (Grafen, 1990). As generalized, costly signaling can play a part not only in mate selection, but in any interaction when underlying traits are difficult to observe (Miller, 2000).

Early modeling in the field showed that punishment can emerge and propagate when one's quality as a future interaction partner is correlated with the cost of publically punishing (Clutton-Brock & Parker, 1995; Gintis, Smith, & Bowles, 2001). This assumption seems particularly reasonable when one considers the degree to which a dominant individual's position allows him to better withstand retaliation. However, under this framing, punishment can sustain cooperative behavior by punishing defectors, but these pressures also favor indiscriminate punishment, making it insufficient to explain the specific targeting of free riders.

Later reputation based models of punishment have broadly fallen into two categories: punishment as a signal of willingness to punish transgressions against the punisher, and punishment as a signal of a disposition supporting local norms. Models of the first case have been primarily focused on second party punishment, meaning that the victim of the transgression themselves punishes the perpetrator. Here we can see that developing a reputation for being willing to punish deters future defection, and therefore reduces the need for punishment (Johnstone & Bshary, 2004). This leads to somewhat paradoxical finding that reputation based strategies lead to both an increased willingness to punish as well as less punishing in equilibrium (McElreath, 2003). Importantly, these models differ from many other explanations for the emergence of punishment in that their proliferation is not due to, even indirectly, the benefits punishment conveys to others in the social group, but rather by simply reducing the chances that the punisher himself is defected against (Dos Santos, Rankin, & Wedekind, 2011). Although limited, experimenters have found evidence consistent with the predictions of the punisher's reputation as a deterrent hypothesis. In a game in which participants could take money from one another as well as punish those who took from them, participants chose to take less from those who previously punished defectors in a cooperative dilemma (Barclay, Submitted).

The other primary thrust of the reputation literature has focused on the relationship between punishment and the punisher's commitment to the violated social norm. When punishing the violation of a social norm, the punisher takes on a cost to express their own underlying endorsement of the norm. If that endorsement also leads to their own conformity with the norm, then punishers would make more reliable partners (Barclay, 2010). Raihani and Bshary (2015) similarly argue that, given that the punishment of

defectors is itself cooperative, punishment can reveal and underlying cooperative disposition. Although attractive in its simplicity, such an argument alone is insufficient to justify an honest costly signal. For such a signal to emerge, there needs to be either a differential cost or differential benefit for the punisher. Otherwise, the signal would be just as effective for a defector to lure cooperative marks, eliminating any informative value.

Some proposals have been put forth to address this shortcoming. Jordan et al. (2016) develop an analytic model, showing that third-party punishment can indeed evolve as a costly signal of norm conformity via partner choice in future interactions. Underlying this model, the researchers assume that there is a correlation between the benefits of punishment one might receive and the benefits of cooperation. The authors note that punishing the perpetrator is beneficial to the victim, in that it deters future harm. As a result, there is some probability of the victim reciprocating the actions of the punisher, either through reward or by punishing someone who exploits the punisher (indeed, experimental evidence shows that third party punishers are rewarded (Railhani & Bshary, 2015)). Importantly, this expected benefit is different for different individuals in the population, who may be more or less likely to interact in the future than others. The critical assumption then becomes that the probability of the future benefit of punishment correlates with the probability of the future benefit of cooperation. This could be due to a variety of reasons, such as how permanent a member of the community the person is or their location in the structure of the population, both of which may affect the probability with which they will repeatedly interact with the victim they punished on behalf of as well as their past partners in cooperative dilemmas.

Another proposal put forward rests on the fact that if you are to punish the violator of a social norm, you need to know that such a violation exists. Research from both the lab and field demonstrates that the domains of cooperation varies dramatically between cultural groups (Poppe, 2005; Cronk, 2007; Goerg & Walkowitz, 2010), even within the same ecology (Henrich & Henrich, 2007). Although sometimes erroneously assumed otherwise, ethnographic and historic studies of foragers show significant levels of fitness relevant ephemeral interactions (Hill, et al., 2011). Incorporating a degree of psychological realism, in such contexts, not all participants may know what norm may apply in a given situation or if any norm applies at all. In this environment, when someone engages in punishment, they are not only advertising their own endorsement of the norm, but also their knowledge that a norm applies in this situation, what norm that is, and that it was violated (Fessler & Haley, 2003). Here again we can see the possibility for a costly signal to arise. Knowing that a norm exists in the local population is a pre-requisite for both punishing a perpetrator as well as conforming to that norm yourself. Whereas people who do not know what norms apply in the local context are at the risk of the extra cost of punishing when no norm applies, and incurring more retribution as the punishment would be seen as unwarranted, those who do know what norm applies can efficiently signal their knowledge by only applying punishment when a norm is applicable. Similarly, those with knowledge of when particular norms are in play can efficiently cooperate conditional on whether the local ecology requires them to, and free-ride when it does not. The relationship between being able to both selectively punish and selectively cooperate in only the right contexts creates the correlation necessary for the emergence of an effective costly signal.

If such a pressure were in play, where punishment indicates knowledge of the relevant norm, we would expect those who punish to be chosen as cooperative partners. Indeed, a preference for punishers as cooperative partners has been demonstrated in multiple experimental contexts (Barclay, 2006; Jordan et al., 2016). This model would also predict that those who are more confident at what norm applies would be willing to pay more to punish, and that this willingness would serve as an even stronger signal. And indeed, experiments show that the more one is willing to pay to punish, the more they are preferred as an interaction partner in a future cooperative interaction (Nelissen, 2008).

Group Selection and Cultural Learning Models of Punishment

Models discussed in the previous section have focused exclusively on interactions within a group. However, selection can occur at multiple nested levels, both within groups as well as between groups. Group selection is a particularly important level of analysis when there is significant between group variation accompanied by within group similarity. In such cases, selection at the group level can be a stronger force than selection at the individual level, favoring the selection for group beneficial traits despite them being individually costly within the group.

A critical insight of group selection modelers rests in the understanding that although both cooperation and the punishing of defectors are cooperative dilemmas, group composition affects their relative cost quite differently (Boyd, Gintis, Bowles, & Richerson, 2003). Cooperating among a group of defectors is not necessarily any more costly than cooperating within a group of cooperators. For example, the cost of contributing to a public good is constant, whether or not others choose to contribute. However, the cost of

punishment is directly linked with the prevalence of cooperators. When few cooperators are present, nearly every interaction would mandate punishment, making it incredibly costly. However, when few defectors are present, compensation is rarely called for, making it particularly cheap. A cooperative group without punishment therefore has a particularly acute within group selection pressure *against* cooperation, meaning that if group selection where to maintain cooperation, the between group pressure to cooperate would need to be particularly large. This can be contrasted with a cooperative group *with* punishment, in which there is minimal selection pressure to defect (as you would be punished) as well as minimal pressure against punishment (as defection is quite rare). This low degree of within group costs means that the between group pressures need not be particularly strong to sustain pro-social punishment, thereby sustaining cooperation.

It is important to point out that even within such dynamics favoring between group selection, cooperation and punishment are not necessarily evolutionarily stable (Boyd et al., 2003). In addition, many models of this type either implicitly or explicitly take this between group selection to be genetic. Genetic group selection requires sufficient genetic between group differences for selection to act upon (Bowles, 2006). These would require significant differences between neighboring communities, maintained through almost entirely endogamous marriage and minimal migration. However, ethnographic evidence suggests behaviorally this is not the case, consistent with the genetic evidence showing insufficient differences between neighboring communities to support genetic group selection (Hill, et al., 2011; Langergraber, et al., 2011).

Cultural group selection models, coupled with an understanding of culture-gene coevolution, address many of these shortcomings. Cultural group selection suggests that the frequency of culturally transmitted components of an individual's phenotype is affected to some degree by the feature's effect on the proliferation of a cultural group (Chudek & Henrich, 2011). Relatedly, culture-gene coevolution proposes that cultural and genetic evolution interact, each creating novel selection pressures for the other. Unlike alleles, variation in cultural norms tends to be between, rather than within, communities (Boyd, Richerson, & Henrich, 2011). And unlike genetic group selection, migration need not undermine, and can in fact bolster, the effects of cultural group selection, providing an opportunity for the migrant to adopt the norms of the local community (Boyd & Richerson, 2009). Culture-gene coevolution also provides a psychologically richer account of genetically acquired cultural learning mechanisms, such as conformity biased learning, which can solve the stabilization problem experienced by genetic group selection models (Henrich & Boyd, 2001).

The culture-gene coevolution account of cooperative norms and their enforcement begins with humans' adaptation to acquire what is known as cumulative culture: information that could not have been acquired by one individual in a single generation. Cumulative culture, as opposed to cultural information that one could develop within a generation, is one of the earliest distinctive elements of psychology so far known to only be present in humans (Boyd & Richerson, 1996). The advent of cumulative culture created a unique genetic selection pressure; those with the cognitive capacity to best acquire this fitness relevant cultural information were better able to survive and reproduce. As genetic evolution pushed the population to be able to obtain greater amounts of cultural information, this

allowed for the development of more complex content, effectively creating a ratcheting effect; with the developing of cultural information selecting for those better able to acquire and store it, and this adaptation allowing for more advanced information, thereby creating an even stronger pressure to acquire it (Tennie, Call, & Tormasello, 2009).

By this account, culture-gene coevolution led to the development of a variety of social learning strategies (Rendell, et al., 2011). These strategies include selectively following those who are successful or prestigious, as well as generally conforming to what those around you are doing (Henrich & Gil-White, 2001; Kendal, Giraldeau, & Laland, 2009). These cultural learning strategies lead to phenotypic assortment: being more similar to those within your group than the population average. Importantly, this leads to a variety of within community stable equilibria. With a high degree of within group homogeneity and between group heterogeneity, cultural group selection serves as an equilibrium selection mechanism, favoring those groups with cooperative norms. Critical for the explanation of sanctioning, these cultural learning strategies such as conformity biased learning, can serve as a sufficient stabilization device to make punishing non-cooperators an evolutionarily stable strategy (Henrich & Boyd, 2001; Guzman, Rodriguez-Sickert, & Rowthorn, 2007).

Differences in social norms across different contexts and communities, as well as others' willingness to punish violations of these norms, creates a novel evolutionary pressure to quickly learn the social norms of a given community. This can lead to novel social learning strategies specifically attuned to responses to social norm violations, such as punishment. Punishment therefore takes on an additional role, not only directly stabilizing norms

within a community with deterrence, but also educating observers as to what is and is not acceptable, giving punishment value as information (Cushman, 2013). In the cultural group selection paradigm, we therefore expect individuals in groups who punish to do better than those who do not, not only because punishment directly stabilizes local cooperative norms, but also because it indirectly stabilizes them by conveying the normative beliefs of the members.

The Evolution of Third Party Compensation of Norm Violation Victims

We propose a unified social norm violation response system, which detects violations of relevant norms, elicits an emotional response of moral outrage, and results in multiple behavioral response patterns, including both the punishment of perpetrators and the compensation of victims. Given that both punishment and compensation are proposed to be the result of this single evolved system, these behaviors must also have been shaped by some of the same underlying selection pressures. In this section, we evaluate what models for the evolution of punishment may also be applicable to compensation, in order to make a novel extension by expanding them to explain the emergence of a social norm violation response system which includes victim compensation.

In order to best describe the proposed evolutionary account of compensation of norm violation victims, it is important to first have a framework to understand social norms. To this end, we use the concepts developed in Bicchieri's (2006) social norms theory. Under this framework, a social norm is a rule of behavior which people prefer to follow, on the condition that they think a sufficient number of other people follow the rule, and that a sufficient number of other people think that they *should* follow the rule. What one

believes other people do is referred to as her empirical expectations, whereas what one believes other people think she *should* do is termed her normative expectations. It is important to note that one's conformity to the social norm is *conditional* on whether one has sufficient empirical and normative expectations, a prediction born out experimentally (Bicchieri & Xiao, 2009).

One common element of both reputation based models as well as cultural group selection models is that the punishment of a violator signals the punisher's belief that one *should* not violate this particular norm. Under these models, punishment therefore serves the function of increasing the normative expectations of observers that the punisher endorses the violated norm. However, simply increasing the normative expectations of observers is not in and of itself an evolutionary explanation. Under reputation-based models, increasing the normative expectations of observers might be evolutionarily relevant by demonstrating that you are a resident of the community and understand the norms of the local ecology, thereby making you a preferable partner in future interaction. Under cultural group selection models, increasing the normative expectations of observers, who have evolved a norm psychology to specifically learn and conform to beliefs about what others think they should do, stabilizes local norms, allowing for between group selection to promote the proliferation of members of groups which stabilize cooperative norms through punishment.

We believe the general claim that punishment serves as an indicator of one's personal beliefs about what people should do, as well as the specific reputational and cultural group selection models outlined above, could also apply to the compensation of norm violation

victims. In the following sections, we describe how those proposals could be expanded to account for compensation, allowing for the development of testable predictions assessed in Chapter 3.

The Norm Broadcasting Hypothesis

Across both reputation and cultural group selection models of punishment, punishment does not only serve a deterrent function, but also functions to convey information. By punishing the violator of a social norm, one reveals his own normative beliefs. This revelation can signal that the punisher is aware of the relevant norms in a particular context and therefore capable of conforming (Fessler & Haley, 2003), or as a direct indicator of endorsement and cooperative intent (Jordan et al., 2016).

Cultural group selection models also use punishment to convey information, albeit for a different underlying rationale. Culture-gene coevolution accounts of norm psychology posit a suite of adaptations for both signaling and receiving signals of local norms (Chudek & Henrich, 2011). Punishment therefore directly stabilizes social norms through deterrence, but also indirectly by conveying the local norms, signals that others are predisposed to follow (Cushman, 2013).

Similar to punishment, compensation could function to signal the third parties normative beliefs concerning the violated norm. Unlike cheap talk, compensation parallels punishment in that it is directly costly to the third party, giving credibility to the signal. Although both punishment and compensation convey information, compensation does *not* have the direct deterrent effect of punishment, which might lead one to wonder how a system would emerge to signal via both compensation and punishment, when punishment

has a clear additional benefit. While compensation lacks a direct deterrent effect, it has unique benefits over punishment, making the two signals complementary. When a norm violation occurs, it is not always clear who committed the violation. If one could only avail themselves to punishment, they would have no opportunity to signal their endorsement of the norm. Assuming that victims are more readily available than perpetrators, or at least available in different situations, compensation has unique benefits. Consistent with this understanding, experiments have shown that when the perpetrator is unavailable for punishment, third parties are more willing to compensate their victim (Chavez & Bicchieri, 2013; Jordan et al., 2016). And although compensation may carry more upfront costs to the third party than punishment (Baron, 2007), it also does not carry punishment's downstream risk of retaliation (Chagnon, 1988; Cinyabuguma et al., 2006; Nikiforakis, 2008; Skarlicki & Folger, 1997). Punishment and compensation can therefore complement one another while serving the same underlying function.

Based on this reasoning, we propose the norm broadcasting hypothesis: *the function of compensating the victim of a social norm violation is to signal the compensator's endorsement of the violated norm*. This hypothesis is not, however, an evolutionary one. To be so, we would need to explain how a third party derives fitness benefits from signaling their endorsement of the violated norm. In the following two sections we provide two candidates for such an explanation, the first based on the reputation models of punishment, and the second based on the cultural group selection accounts. It is important to note that neither of these two proposals are mutually exclusive. It could be the case that both reputational concerns and cultural group selection jointly explain the

compensation of norm violation victims, or either individually. Between these possible combinations, we are agnostic.

The Reputation Signaling Hypothesis

Reputation based explanations for punishment broadly fell into two camps. The first was that individuals developed a reputation as a punisher in order to deter future defection against themselves (Johnstone & Bshary, 2004; Dos Santos et al., 2011). The second was that people punish to build a reputation as someone who is embedded in the local social community and knows and endorses the local social norms, thereby improving their perceived quality as a future partner (Jordan et al., 2016; Fessler & Haley, 2003). As compensation does not provide a direct deterrent effect, it is not obvious how the models in which building a reputation as a punisher deters future defections against you could be extended to account for compensation. One possibility points to the fact that when the perpetrator is unknown, compensation serves as a substitute signal. This would mean that while compensation is not itself a deterrent, if it is correlated with a willingness to punish, it could indirectly build one's reputation as a punisher. Experimental results do suggest that a willingness to punish is correlated with a willingness to compensate (Jordan et al., 2016), consistent with this speculative hypothesis.

However, we find the models under which one punishes in order to indicate their membership in a community and knowledge of local social norms to more naturally lend themselves to also explaining victim compensation. Jordan et al. (2016) showed how punishment could evolve as a signal of one's trustworthiness, due to the correlation between future benefit of punishment and the future benefit of cooperation. This is

because the benefit of each is partially determined by how transient they are in a population. When you reside within a community permanently, you are more likely to benefit from having punished on someone else's behalf through the reciprocity of them being willing to punish on your behalf, thereby deterring those who might otherwise exploit you. Similarly, if you are a permanent resident, you would also be more likely to benefit from repeated cooperative interaction. This correlation in benefits allows punishment to be a costly signal of cooperative intent.

This same signaling argument can be expanded to account for compensation. If you compensate on someone's behalf, you are more likely to benefit from their reciprocated compensation in the future if you reside within that community. Similarly, if you cooperate with someone, you are more likely to benefit from their continued willingness to interact with you if you are in the same community. Therefore the relative benefits of compensation are correlated with those of cooperation, leading to compensation serving as a truthful signal of willingness to cooperate. Given that, we would expect others who are looking for partners in a cooperative dilemma to be attuned to candidates' reputation for compensation.

In addition to the probability of future interaction, punishment can signal one's knowledge of the norms within a local community (Fessler & Haley, 2003). This explanation hinges on the fact that social norms for the same situation can differ dramatically from one community to another, even within the same ecology (Cronk, 2007; Goerg & Walkowitz, 2010; Henrich & Henrich, 2007; Poppe, 2005). It would be incredibly individually costly to cooperate in every situation where a cooperative norm *could* apply.

Similarly, it would be overwhelmingly costly to punish free-riders in every such situation. However, if one knows the local norms, one can cooperate and punish at much lower frequency, and even at higher individual cost, while not risking needless cooperation or being punished. This creates the correlation necessary for a truthful costly signal: cooperating is relatively less costly for those who know the local norms, as they need only to cooperate in the specific situations dictated by the norms of that community. Similarly, punishment is less costly for those who know the local norms, as they need only do so when a norm of that specific community is violated. This leads to the punishment of local norm violations functioning as costly signal of conformity to local norms.

This logic can be directly expanded to include compensation as well as punishment. Compensating the victims of all free-riders is far more costly than only compensating those who were the victim of a social norm violation, as deemed by the local community. This results in the same correlation as described above for punishment, between the relative cost of compensation and the relative cost of punishment. This correlation allows for compensation of the victims of local norm violations to serve as a costly signal of conformity to those local norms. From this logic, we expect a downstream consequence to be that those in search of a cooperative partner to be attuned to candidates' compensatory reputation. Taking these possible accounts for the emergence of compensatory behavior, we propose the reputation signaling hypothesis: *compensators signal their status as a member of the community and their knowledge and support for local norms, leading observers to prefer interacting with the compensator in the future*.

The Norm Stabilization Hypothesis

Group selection accounts allow for selection not only between individuals, but also between groups. Looking at both these levels, we can see that within a group, if there is a high willingness to punish, this entails low levels of defection, meaning that punishment will rarely be used. This results in the within group selection pressure *against* punishment being weak. This can be contrasted with the strong between group pressure when punishment has stabilized cooperation in one group and not another, making between group selection pressures particularly relevant in the domain of cooperation and punishment. Culture-gene coevolution models come from the realization that, due to human's unique ability to acquire cumulative culture, solely genetic explanations may be insufficient to properly understand the dynamics of human cooperation, including third-party punishment. By their account, the emergence of cumulative culture created a genetic selection pressure to effectively acquire that information, which resulted in a suite of cultural learning strategies such as conformity and imitation (Rendell, et al., 2011). The bias in these cultural learning strategies can be sufficient to counteract the weak selection pressure against punishment within groups (Henrich & Boyd, 2001).

Movement between and interaction with other cultural groups, as is prevalent in hunter gatherer cultures today (Hill, et al., 2011), creates a strong pressure to learn what social norms were endorsed by the local community to avoid the previously established punishment. Punishment can therefore serve not only as a direct stabilization device of local norms, but also as a teaching tool (Cushman, 2013), counteracting the normative expectation reducing effect of having observed someone violate a local norm, thereby stabilizing the norm at precisely the time when it might otherwise be undermined.

Unlike punishment, victim compensation does not provide a direct deterrent effect. However, if people have developed a suite of cognitive tools specifically to acquire local social norms, then compensation could plausibly serve a similar teaching function. Just like punishment, compensation is costly. Therefore there would only be selection for one to compensate (or punish) if there is in fact a social norm in the community that the compensator (or punisher) would benefit from the stabilization of. As the benefit of compensation is correlated with there actually being a relevant social norm, compensation can function as an effective costly signal of local norms. Given the previously discussed evolved psychology for acquiring and complying with local norms, we would then expect that observing compensation would induce observers to comply, thereby stabilizing the violated norm. From this logic we develop the norm stabilization hypothesis: *Compensators signal their normative belief that one should not violate the norm, inducing observers to comply with the norm, increasing its stability*.

The norm broadcasting, reputation signaling, and norm stabilization hypotheses are novel accounts for the functional underpinnings of the compensation of the victims of norm violations. As we based these accounts on expanding evolutionary models of punishment of norm violators, they allow us to start to speak of a unified evolutionary account of third party response to the violation of social norms, consistent with our understanding of the proximate emotional motivators established in Chapter 1. In Chapter 3, we take each of these three hypotheses and derive previously untested predictions. We then experimentally test those predictions to build an empirical understanding of what selection pressured may have resulted in this norm violation response system.

CHAPTER 3

Chapter 1 provided evidence for a finer-grained picture of the emotional motivations for compensation. While previous research had taken empathic concern broadly as the primary motivator for compensatory helping behavior, our work showed that, in the case of a social norm violation, moral outrage drove third parties to compensate.

As emotional states can taken as indicators of the underlying psychological mechanism at play (Fessler & Haley, 2003; Haidt, 2003), third party compensation being motivated by empathic concern in some contexts, but by moral outrage in the case of norm violations, would therefore suggest that these behaviors are driven by distinct underlying mechanisms. The disassociation between the mechanism underlying the compensation of norm violation victims and compensation in other contexts presents an evolutionary puzzle. If both forms of compensation arose from the same pressures, we would not expect distinct psychological mechanisms. Therefore, the explanation for empathic compensation, centering on the indirect reciprocity benefits of efficient transfer of goods to those in need (ex. Nettle et al., 2011), cannot provide the evolutionary rationale for the morally outraged compensation of the victim of a social norm violation. It is this puzzle which motivates this chapter: what led to the evolution of third party compensation of norm violation victims?

Whereas the anger underlying the compensation of victims of norm violations is different than the emotions motivating compensation in other contexts, it coincides with the emotional antecedents of punishment (Fehr & Fischbacher, 2004; Jordan et al., 2017; Nelissen & Zeelenberg, 2009). This suggests that the *compensation* of norm violation

victims may be more closely related to the *punishment* of norm violation perpetrators than it is to other types of compensation. Given these similarities at the emotional level, we might also look to the similarities at the cognitive level. Just as punishment has been shown to be driven by the desire to right a moral wrong (Carlsmith, 2006), the compensation of a victim can achieve the same fairness restoring goal, but in the domain of the victim rather than the perpetrator.

And although compensation cannot provide the same deterrent function as punishment, it *could* plausibly serve similarly as a signaling device, revealing the compensator's endorsement of the norm. Compensation rather than punishment has the added benefit of not incurring the risk of retaliation from the punished perpetrator, a fear justified in both the lab and field (Chagnon, 1988; Cinyabuguma et al., 2006; Nikiforakis, 2008; Skarlicki & Folger, 1997). In addition, compensating victims, rather than rewarding compliers, shares the positive trait with punishment of, when the norm is stable and compliance is high, being unnecessary and therefore low cost (Oliver, 1980).

Based on the similarities in both motivation and plausible cognitive rationale between punishment and victim compensation, we propose the norm broadcasting hypothesis: *the function of compensating the victim of a social norm violation is to signal the compensator's endorsement of the violated norm*. This logic parallels that of some evolutionary models of punishment, which suggest that punishment may serve as a signal to both the punished individual as well as those observing (Barclay, 2006; Cushman, 2013; Jordan et al., 2016; Kurzban et al., 2007). However, from an evolutionary prospective, this logic alone is insufficient, as it does not explain how the compensator derives net fitness benefits from

their costly signaling of their endorsement of the violated norm, a necessary condition for such disposition to have evolved.

For this, we propose two candidate hypotheses, between which we are agnostic, both of which could be simultaneously contributing to the emergence of the compensation of norm violation victims. We first have the reputation signaling hypothesis: *compensators signal their status as a member of the community and their knowledge and support for local norms, leading observers to prefer interacting with the compensator in the future*. This proposal closely mirrors the argument found in indirect reciprocity models of third-party punishment, which propose that punishing builds one's reputation as a quality partner (Barclay, 2006; Jordan et al., 2016).

Our second candidate is the norm stabilization hypothesis: *compensators signal their normative belief that one should not violate the norm, inducing observers to comply with the norm, increasing its stability.* Assuming the norm in question is a cooperative norm, the stabilization of the norm benefits the compensator by increasing the probability of future interactions being cooperative. The norm stabilization hypothesis hinges on our understanding of the psychological determinants of norm compliance. Previous work demonstrates that individuals conform to a norm only if they believe that a sufficient number of other people conform (empirical expectations) and that they believe that a sufficient amount of others think they *ought* to conform (normative expectations) (Bicchieri & Chavez, 2010; Bicchieri & Xiao, 2009). Previous studies have shown that when people witness the violation of a social norm, it decreases their relevant normative expectations, leading the observer to be more willing to violate the norm themselves

(Bicchieri & Xiao, 2009). The norm stabilization hypothesis therefore posits that a third party can send a costly signal that they do in fact endorse the social norm, thereby providing reinforcement of the norm in the eyes of observers. Similar proposals have been made to explain the emergence of third-party punishment, in which punishment is taken as a teaching tool, informing both the punished and those observing of what is acceptable behavior (Cushman, 2013).

There are currently no other evolutionary models designed to explain a compensation system which is engaged specifically in the case of a norm violation. Therefore, we did not design the studies in this chapter to pit the predictions of our hypothesis against those in the literature. Instead, we attempted to test the norm broadcasting, reputation signaling, and norm stabilization hypotheses by determining novel downstream predictions under each, and testing those predictions.

In Study 4, we investigated the general norm broadcasting hypothesis, which posits that the function of compensation is to serve as a costly signal of the compensator's endorsement of the norm. As in any signaling system, the efficiency of such a system is entirely dependent on others receiving that signal. We therefore posited that *if* victim compensation did arise in order to signal the compensator's endorsement of the norm, one's willingness to compensate would be moderated by the number of people who observe that compensation. We tested this prediction by having participants take part in a third party trust game. In the role of the third party, participants decided whether to, in the case of the investor receiving no money due to the trustees' violation of a reciprocity norm, pay $5 of their own endowment to restore the investor to their original $15

endowment. All participants were assigned to either the public or private condition, where their choice was either revealed to a group of disinterested observers or kept private. Under the norm broadcasting hypothesis, we expected more participants to be willing to compensate when their choice was observed. Although no such effect has previously been investigated in the domain of victim compensation, the mirrored effect in punishment has been found, in which being observed increases third party's propensity to engage in the costly punishment of norm violators (Kurzban et al., 2007).

Study 5 investigated the predictions of the norm stabilization hypothesis. The norm stabilization hypothesis suggests that the signaling of one's endorsement of the violated norm through compensation functions to mitigate the norm undermining effect of witnessing a violation, increasing the probability that the norm is maintained. This hypothesis relies on the logic that if one witnesses the compensation, this increases that person's normative expectation that other people believe one ought to follow the norm, which increases their own compliance. In Study 5, we tested the prediction of the norm stabilization hypothesis that observing compensation increases the observers' propensity to conform to the norm. Participants were assigned to the role of a trustee in a trust game. Before making their choice as a trustee, participants were shown summary statistics of other participants in a third party trust game. Participants in the High Compensation condition were shown high levels of compensation by third parties, whereas participants in the Low Compensation condition were shown low levels of compensation. Under the norm stabilization hypothesis, we predicted that witnessing high levels of compensation would lead trustees to be more likely to comply with the reciprocity norm, and that this effect would be mediated by their normative expectation. Similar effects have previously

been shown in the punishment domain. There we have seen that when one is punished for violating a norm they are more likely to cooperate (Fehr & Gachter, 2000), even when punishment is no longer present (Stagnaro, Arechar, & Rand, 2017), and that having witnessed someone else be punished can have an even stronger effect than having been punished yourself (Barr, 2001).

Finally, in Study 6, we test predictions of the reputation signaling hypothesis. Under this hypothesis, people use compensation as a costly signal of their membership in the local community and knowledge of the relevant local norm in order to increase their attractiveness as a partner in future interactions where the same norm applies. In order to assess the reputation signaling hypothesis, we therefore chose to test the prediction that observers do in fact prefer to interact with those who previously compensated. Participants in this study were told that they were to assume the role of an investor in a trust game. They were then shown a set of possible trustees, all who had acted pro-socially in a previous interaction, but only one of which had compensated the victim in a third party trust game. Under the reputation signaling hypothesis, we predicted that people would be more likely to select the compensator. Similar effects to that proposed here have been observed in the punishment domain, showing people to prefer punishers over non-punishers in future interactions (Barclay, 2006; Jordan et al., 2016).

Study 4

A key principle of the general signaling hypothesis is that compensation functions to transmit information to those observing the act. This leads to the prediction that possible compensators should be attuned to the degree to which their compensation is observable

(and therefore a more or less effective signal). Specifically, this leads to the hypothesis that third parties should be more willing to compensate the victims of a norm violation when that choice to compensate is made publicly. This study tested that proposition.

Participants took part in a trust game with third party compensation. Using the strategy method, each participant fully described what they would do in each role of the game, including whether or not they would compensate an investor who lost their endowment due to the norm violation of the trustee. Each participant was then randomly assigned to one of the three roles, and their decision was carried out. Critically, participants were divided into either a Public or Private condition. In the Public condition, all participants were required to state their decision aloud, whereas in the Private condition, their choices were left undisclosed. We predicted that, consistent with the norm signaling hypothesis, that third parties would compensate significantly more if they knew they would make their decision publicly.

Method

We recruited 153 participants (45 men, mean age of 21) from the PLEEP subject pool to participate in this study. The PLEEP subject pool consists of student and staff members of the University of Pennsylvania who have enrolled to receive notification of studies occurring on campus.

Participants were invited to the lab in groups of 10 to 16 across 12 sessions. All participants were paid a show-up fee of $5. Each participant then read a description of a third party trust game, similar to that used in the norm violation situation in Study 1. The investor was endowed with an additional $10, the trustee with $0, and the third party with $15. The

investor could choose to either keep the $10 or transfer the entire amount to the trustee. If the investor chose to transfer, the experimenter tripled the amount to $30. The trustee could then chose to either keep the entire $30, or transfer $15 of the $30 back to the investor. Finally, if the trustee chose to keep the entire $30, the third party could choose to either keep their entire $15, or to transfer $5 to the investor. If the third party chose to transfer $5 to the investor, the experimenter doubled this amount to $10. The options available in each role were common knowledge to all participants.

Before knowing their role in the interaction, each participant chose what they would do if assigned to each of the three roles (i.e. each participant said, as the investor, whether they would transfer the money; as the trustee, whether they would return the money if transferred to them; and as the third party, whether they would pay to compensate the investor if the money was not returned to them). This commitment was final, and could not be changed later in the session.

The experimenter informed participants that there were multiple sessions of the experiment, and that each *session* was assigned to a particular role (i.e. the entire session would be either investors, trustees, or third parties). In addition, each session was randomly assigned to be either Public or Private. In the Public condition, subjects were informed that after making their decisions and being assigned to a role, they would verbally announce their choice to the experimenter and other members of the session. Previous studies have used this technique successfully the manipulate audience effects in behavioral games (Chavez & Bicchieri, 2013; Kurzban et al., 2007).

We chose the method of soliciting choices for all roles from each participant and then assigning the entire session to a particular role for two reasons. First, soliciting choices from each participant for all roles made efficient use of subject responses, as we only analyzed the responses for the role of the third party. To reduce spill over between responses across roles, we solicited responses for the role of the third party *before* the two other roles. Second, assigning the entire session to a particular role rather than assigning each individual was designed to reduce the demand effect of having to publicly declare one's choice *to the affected individual*. The signaling hypotheses does not require that those observing be in any way involved in the particular exchange, so this method provided a stronger test, as all participants in any given session were by design not interacting with each other, as they were all assigned to the same role. After all sessions were complete, participants were matched with one another, and notified that they could pick up any additional earnings.

Results

All participants correctly responded to the four comprehension questions and were included in the analysis.

Proportions of participants who chose to compensate for the Public and Private conditions can be found in Figure 5. We found that a significantly higher percentage of participants in the Public condition (76.7%) than in the Private condition (38.8%) chose to compensate $X^2(1, N=153)=20.9$, $p<.001$.

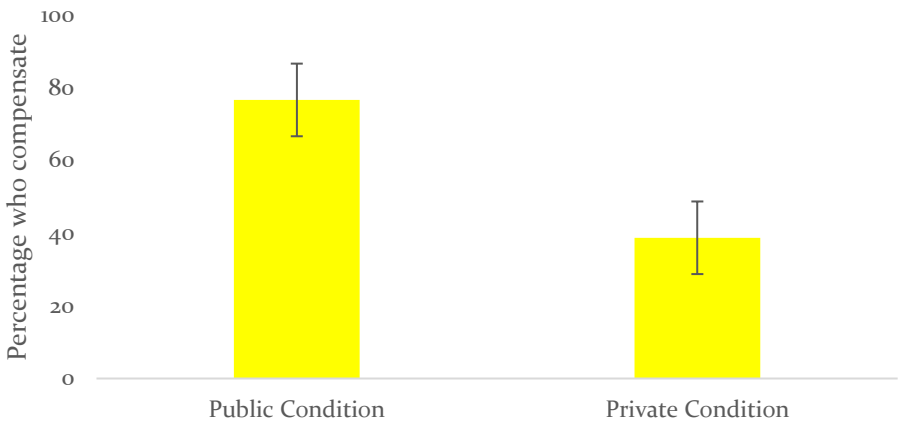

Figure 5. Percentage of participants who choose to pay $5 to compensate investor in the public versus private condition, error bars as 95% confidence intervals

Discussion

The general norm broadcasting hypothesis sees compensation as functioning to *broadcast* the beliefs of the compensator. The effectiveness of broadcasting is proportional to the size of the audience. We therefore expected, if compensation functions as suggested, for it to be attuned to whether or not an audience is present. Here we find support for this prediction, with participants randomly assigned to the Public condition compensating significantly more than those assigned to the Private condition.

Study 5

This study, along with Study 6, begin to disentangle the alternative justifications for the norm broadcasting hypothesis: the norm stabilization hypothesis and the reputation signaling hypothesis. In this study, we focused on predictions made by the norm stabilization hypothesis, which argues that compensation broadcasts one's endorsement of a norm, which in turn limits the destabilizing effect of a norm violation, and increases

relative conformity. Bicchieri's (2006) theory of social norms proposes that one condition for social norm conformity is that a sufficient number of relevant others think that one should conform. Previous work shows people to in fact have such conditional preferences for norm conformity (Bicchieri & Chavez, 2010; Bicchieri & Xiao, 2009). We propose that compensation can broadcast those expectations, thereby serving as a form of "psychological deterrence" due to people's conditional preference for norm conformity.[2]

In order to test the prediction of the norm stabilization hypothesis, participants were shown summary statistics of other peoples' past behavior in a third party trust game. They were told the proportion of investors who invested, the proportion of trustees who returned the funds if invested, and the proportion of third parties who were willing to pay to restore the investor to their original endowment if the trustee kept the entire transfer. Participants were assigned to either the High or Low Compensation condition. In both conditions, subjects were shown summary statistics of select cases, showing moderately high rates of investment and return. In the Low Compensation condition, participants were shown the summary statistics of select cases in which a low proportion of third parties compensated, whereas participants in the High compensation condition were shown summary statistics from select cases in which a high proportion of third parties compensated. Participants then took part in a trust game themselves, with *no* third party. This study tested the prediction of the norm stabilization hypothesis that trustees who

[2] It is important to note here the similarities and differences between punishment and compensation. Unlike compensation, punishment has an immediate *materially* deterrent effect. However, punishment and compensation are similar in that they both signal the expectations of the third party. This allows the third party to honor the norm in a costly manner, thereby signaling their own endorsement of the norm with minimal threat of reprisal.

observe high rates of compensation would be more likely to conform to the norm by returning half the funds back to the investor. As a secondary prediction, the norm stabilization hypothesis also predicts that it is through the path of increasing normative expectations, the belief that other people think you should conform to the norm, that high levels of compensation lead to greater norm conformity.

Method

We recruited 1098 participants (543 men, mean age of 36) from the AMT platform to participate in this study.

We divided this study into two phases. In Phase 1, participants read a description of a simplified trust game with third party compensation. The investor received an initial endowment of $0.50 the trustee $0.00, and the third party $0.75. If the investor chose to transfer to the trustee, their transfer of $0.50 was tripled to $1.50. The trustee then chooses whether to keep the entire $1.50 or to send half ($0.75) back to the investor. If the trustee chose to keep the entire sum, the third party had the option to pay $0.25 to restore the investor to $0.50. All options for all roles were common knowledge. Participants were then randomly assigned to one of these three roles, and made their choice for that role. We ran participants in Phase 1 until we had a proper mix of example participants to use in Phase 2. This meant running participants until at least 1 investor chose to invest, at least 1 investor chose not to invest, at least 3 trustees chose to return the money, at least 1 trustee chose not to return the money, at least 9 third parties chose to compensate, and at least 9 chose not to compensate. After these conditions were met, we moved on to Phase 2.

In Phase 2, participants were shown the third party trust game described above. They were then shown what was described to them as "a sample of the results" of previous participants. How that sample was constructed was not described. The sample was taken from Phase 1, chosen such that every subject was shown that 50% of investors chose to transfer, and 75% of trustees chose to return. However, the proportion of third parties who chose to compensate was manipulated through a selection of cases from Phase 1. Specifically, in the high compensation condition, subjects were shown a sample in which 90% of third parties chose to compensate the investor, whereas in the Low Compensation condition subjects were shown a sample in which only 10% of subjects chose to compensate the investor.[3] All three statistics, rather than only the proportion of third parties who compensated, were shown in order to reduce any demand effect and provide a particularly strong test of the hypothesis.

After each participant viewed the summary statistics described above, they were then given the opportunity to play a trust game themselves, similar to that played in Phase 1, but with no third party. Before telling participants which role they would take in the game, we employed the strategy method so that all participants committed to their choices for both roles before being told which role they would actually take. In order to reduce spillover effects, all participants chose what they would do as the recipient first.

[3] Previous studies have used similar methods of deliberate rather than random selection to manipulate information through the use of an ambiguous prompt such that they can deliver consistent stimuli while maintaining non-deception (Bicchieri & Xiao, 2009; Charness, Naef, & Sontuoso, 2016).

After making their choices for the interaction, all participants were asked what their personal normative belief and normative expectations were concerning the actions of the trustee in their interaction. We measured personal normative belief by asking participants if they "thought it was wrong for Person B to keep the $1.50". We measured normative expectation, one's belief about the personal normative beliefs of others (Bicchieri, 2006), by asking "Out of 10 participants in this study, how many do you think said it was wrong for Person B to keep $1.50".

Results

657 participants (60%) correctly responded to the four comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

Our primary interest was whether observing high levels of compensation leads to norm conformity. The percentage of respondents who chose to conform to the norm by returning half the transfer to the investor in the High Compensation versus Low Compensation condition can be found in Figure 6. We found that significantly more participants who observed the high compensation (76.9%) than those who observed low compensation (70.1%) chose to conform to the social norm by returning half the transfer, X^2 (1, N=657) = 3.92, p = .048.

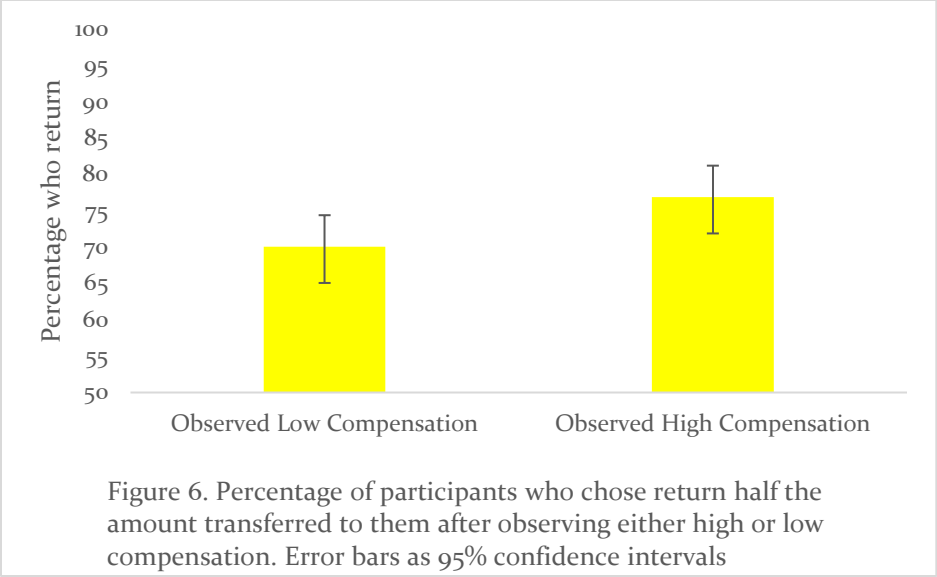

Figure 6. Percentage of participants who chose return half the amount transferred to them after observing either high or low compensation. Error bars as 95% confidence intervals

Those who observed low compensation also reported lower levels of normative expectation (M=6.48) than those who observed high compensation (M=6.84), t(642)= 2.25, p=.025. We therefore conducted a mediation analysis, shown in Figure 7. We found that normative expectation mediated the effect of observed level of compensation on willingness to conform to the norm by returning half the transfer, B=.13, 95% CI=.01 to .25.

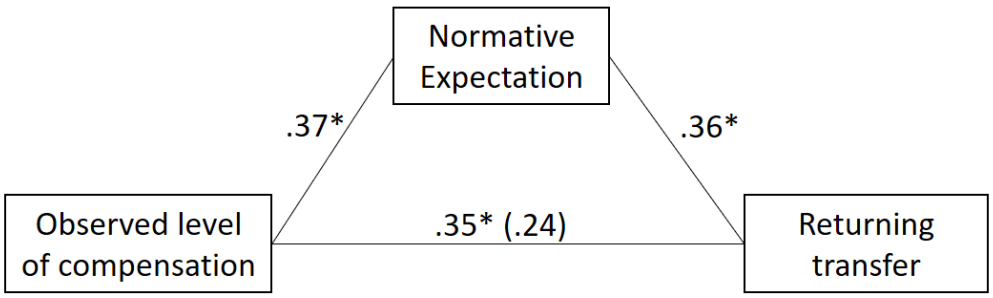

Figure 7. Mediation model for the relationship between observed level of compensation and returning the transferred amount, as mediated by normative expectation. Relationship between observed level of compensation and returning the transferred amount, controlling for normative expectation, in parentheses. *p<.05

Discussion

The norm stabilization hypothesis proposes that compensation functions to maintain social norms by broadcasting the compensator's endorsement of the violated norm. This hypothesis draws on Bicchieri's (2006) theory of social norms, which proposes that one of the conditions for conforming to a social norm is a sufficient number of relevant others believing that you ought to. This concept is supported by previous work showing that whether or not one conforms to a norm is conditional on whether or not they believe that others endorse that norm (Bicchieri & Chavez, 2010; Bicchieri & Xiao, 2009). Therefore, if compensation can signal one's endorsement, then that can contribute to other's conformity, thereby stabilizing and sustaining the norm.

This study tested two components of the norm stabilization hypothesis. First, we tested the effect of observing compensation on one's normative expectation that others thought that it was *wrong* to not return the funds. Here we found that observing compensation did indeed increase observer's normative expectations. Second, we tested whether observing compensation drove one to conform to the norm when no third party was present. Here we found that observing compensation lead participants to conform to the norm significantly more frequently. Bicchieri's (2006) social norm theory provides a conceptual linkage between these two findings, suggesting that observing compensation effects norm conformity via the increase in normative expectation. In testing this, we found that normative expectation significantly mediated the effect of observing compensation on the choice to compensate. This suggests the causal pathway, as predicted by the norm stabilization hypothesis and social norms theory, that observing compensation leads to higher normative expectation, which in turn leads to higher levels of compensation.

It should be acknowledged that the size of the primary effect, an increase of 6.8%, from 70.1% to 76.9%, is relatively small. Given that, it is also important to point out the components of the study design that provides a particularly strong test of the hypothesis, but would be expected to reduce the size of the effect. Past work has pointed to the importance of attentional focusing on normative and empirical expectations in norm conformity (Cialdini, Reno, & Kallgren, 1990; Kallgren, Reno, & Cialdini, 2000). In order to provide a particularly strong test of the norm stabilization hypothesis, we wanted to avoid any level of artificially high focus on the level of compensation. We therefore included both the proportions of investors who invested and trustees who returned. The inclusion of the proportion of trustees who returned is a particularly strong focal signal of both normative expectation (revealing how acceptable the trustees think it is to not return the money) and empirical expectation (directly informing the participant how often people return money as trustees). Past research has shown that when empirical and normative expectations are in conflict, people normally act in accordance with their empirical expectations (Bicchieri & Xiao, 2009). Given this, our manipulation of the level of compensation runs into the restricting force of strong empirical expectation and normative expectation signals from the behaviors of the trustees. We therefore see overcoming these countervailing effects as a particularly robust test of the hypothesis, and find the smaller effect size reasonable.

Study 6

This study serves as a compliment to Study 5, helping disentangle the norm stabilization hypothesis and reputation signaling hypothesis by testing predictions specific to the reputation signaling hypothesis. This hypothesis posits that individuals compensate in

order to inform observers that they know and endorse the norm, and would therefore be worthwhile partners in future interactions. For this to have been selected for, observers must in fact prefer to interact with compensators. Study 6 aims to test this prediction.

In this study, subjects were told that they will soon play a trust game. They were told that they have been selected to play as the investor, and will have some say in who will be the trustee. They were then shown three candidates who could possibly be the trustee in their interaction. For each possible trustee, they were shown information about the possible trustee's past behavior in a previous interaction. The first possible trustee compensated an investor in a trust game. The second possible trustee gave an even split in a dictator game. Finally, the third possible trustee cooperated in a prisoner's dilemma. Through an incentive compatible elicitation, participants then ranked the three possible trustees. As compensation is taken to indicate a disposition to conform to the norm, the reputation signaling hypothesis predicted that investors would prefer the trust game compensator over the two alternative trustees.

Method

We recruited 452 participants (215 men, mean age of 36) from the AMT platform to participate in this study.

We divided participants into two phases. In the first phase, participants were assigned to play either a trust game with third party compensation, a dictator game, or a prisoner's dilemma. Those assigned to the trust game with third party compensation participated in the interaction as described in Phase 1 of Study 5. Those assigned to the dictator game were assigned to one of two roles, either dictator or recipient (labeled as 1 or 2 in the

interaction). The dictator was given a starting allocation of $0.75 and the recipient $0.00. The dictator was then given the opportunity to transfer $0.25 of their allocation to the recipient, which would be doubled by the experimenter. Those in the prisoner's dilemma game participated in a two party interaction in which each party simultaneously chose to either cooperate or defect (labeled UP or DOWN in the interaction). If both parties chose to cooperate, each received $0.50. If both parties chose to defect, each received $0.25. If one person choose to cooperate, while the other chose to defect, the person who chose to cooperate would get $0.00, whereas the person who chose to defect would get $0.75. Importantly, although the context surrounding the decision of the third party in the trust game, the dictator, and the participant in the prisoner's dilemma were different, the fundamental choice they made was the same. In each case, the participant chose whether or not to increase the amount received by another participant by $0.50 at a cost of $0.25.

After making their choices, the trust game, as in Phase 2 of Study 5, was described to all participants. They were then told that they may be assigned to the role of trustee in that interaction in the future. They were then asked, if they were given the opportunity to act as the trustee in the trust game and the investor transferred to them, whether or not they would choose to transfer back half their sum to the investor. We ran participants in the Phase 1 trust game with third party compensation until a third party had the opportunity to compensate and chose to do so. We ran participants in the dictator game until a dictator chose to transfer to the recipient. We ran participants in the prisoner's dilemma until a pairing both chose to cooperate. The third party from the final trust game, the dictator from the final dictator game, and one of the cooperators from the final prisoner's dilemma were then selected as possible trustees for Phase 2.

In Phase 2, participants began the study by reading the instructions for the trust game, as described in Phase 2 of Study 5. They were then told that they had been assigned to the role of the investor, and now had some say in who would be the trustee in their interaction. Participants were then shown the list of the three possible trustees. With each trustee, they were given a description of what that person did in the game they participated in in Phase 1 (i.e. compensating in the third party trust game, giving in the dictator game, or cooperating in the prisoner's dilemma). They were then asked to rank these three possible trustees in order of who they would most like to serve as the trustee in their interaction. They were told that the person they selected first would have a 60% chance of being their trustee, the second person a 30% chance, and the third person a 10% chance. After making their selection, they were paired and asked whether they would like to transfer their endowment to the trustee they were paired with.

Results

404 participants (89%) correctly responded to the four comprehension questions. To ensure data quality, we only analyzed the responses from these participants.

In order to protect against an artificially inflated alpha level, we first tested whether the distribution of participant's top choice of partner was different to that expected by chance. We found that the distribution significantly diverged from the one third selection of each possible partner, as expected by chance, $X^2(2, N=404)=6.27$, $p=.043$.

Having confirmed in the omnibus test that the selection of first choice partners significantly diverged from chance, we then proceeded to assess the selection of each of the three possible

partners as their first choice, relative to the expected one-third selection under chance, can be found in Figure 8. Here we see that participants selected the kind dictator as their first choice 28.5% of the time, significantly *less* than expected by chance $X^2(1, N=404)=4.09$, $p=.043$. We also found that participants selected the cooperator in the prisoner's dilemma 38.6% of the time, significantly more than expected by chance $X^2(1, N=404)=4.83$, $p=.028$. However, we did not find any difference between the 32.9% who chose the compensator in the trust game and the expected rate under random chance, $X^2(1, N=404)=.015$, $p=.90$.

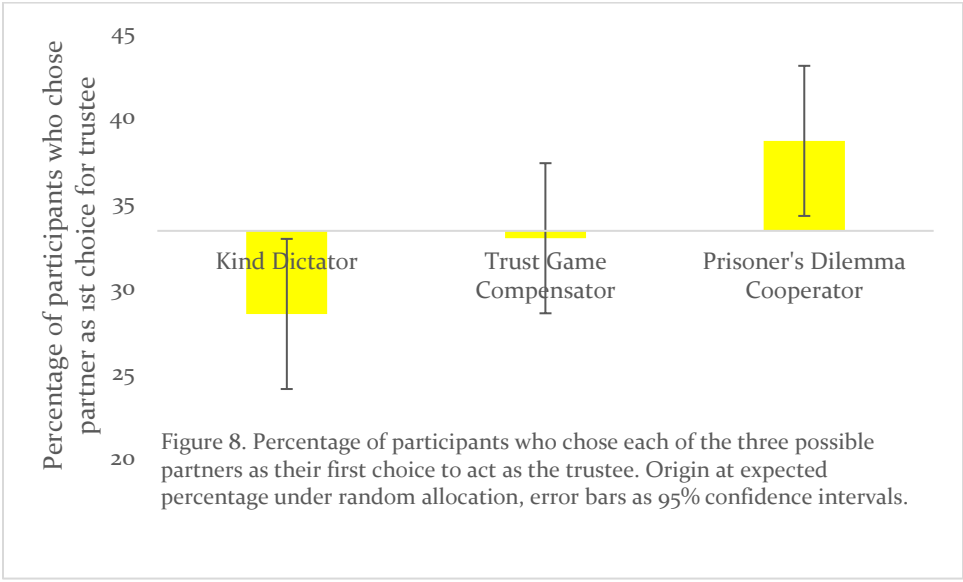

Figure 8. Percentage of participants who chose each of the three possible partners as their first choice to act as the trustee. Origin at expected percentage under random allocation, error bars as 95% confidence intervals.

Discussion

The reputation signaling hypothesis posits that third parties compensate to send a costly signal of their own knowledge and endorsement of a norm, in order to show themselves to be worthwhile partners in future interactions governed by the norm. This mechanism could only have evolved if possible future interaction partners were *attuned* to whether or not a possible interaction partner was a compensator when selecting with whom to

interact. In order to test the reputation signaling hypothesis, we therefore tested whether observers of compensatory behavior did in fact prefer compensators as future interaction partners.

We compared the selection of compensators to the selection of two other possible participants. These other participants, either a fair allocator in a dictator game, or a cooperator in a prisoner's dilemma were different in circumstance. The even allocation in the dictator game suggested that the dictator might be generally prosocial, but such behavior does not represent conformity with a norm (Bicchieri, 2006). Although the cooperator in the prisoner's dilemma did signal their willingness to conform, it was to a cooperation norm rather than the reciprocity norm present in the trust game. Importantly, across all three possible trustees, each one made the choice to increase the amount another participant received by $0.50 at a $0.25 cost to themselves, making the key difference between the candidates the contextual framing of their acts.

Our results showed that the possible trustee's past behavior in these previous interaction did have a significant impact in the selection of a trustee. However, we did not find support for the reputation signaling hypothesis. We observed the lowest rates of selection of the dictator who gave an even allocation, and the highest rates of selection of the prisoner's dilemma cooperator, with the trust game compensator falling in the middle.

Although inconsistent with our formulation of the reputation signaling hypothesis, the results we observed still show an interesting pattern of preferences. The even allocating dictator, which we considered to not broadcast any norm endorsing content but rather just general pro-sociality, was the least selected option, with the two norm broadcasting

possible trustees selected more frequently. This suggests that those selecting a partner may very well be attuned to broadcasting in some fashion. We did not predict that the prisoner's dilemma cooperator would be most frequently selected. We do see a possible explanation for this, which would be explorable in future research. It may be the case that observing norm conformity, regardless of the particular norm, people focus on the fact that the individual conformed, attending less to the particular norm they conformed to. From and evolutionary prospective, this focusing would be justified if it were the case that different individuals have different *general* dispositions towards norm conformity, leading to their conformity to one norm being diagnostic of their conformity to another. This claim is empirically supported, with individuals who conform to pro-social norms in one behavioral game being much more likely to conform in other contexts, even when conducted months apart (Yamagishi, et al., 2013).

In the study in question, it may the case that participants focused on the norm conforming behavior, despite being conformity to a cooperation rather than reciprocation norm, because conformity generally is more diagnostic than compensation in the specifically relevant context. This explanation would therefore be consistent with a broadened understanding of reputation signaling, where your reputation in the domain of one norm is linked to your reputation in other norm contexts. This suggested explanation is contingent on a low degree of *within* individual heterogeneity in norm conformity across a variety of norms, as compared to a relatively high degree of *between* individual heterogeneity in general disposition to follow norms, an empirical question yet to be addressed.

General Discussion

In this chapter, we investigated the three hypotheses developed in Chapter 2: the norm broadcasting hypothesis, the reputation signaling hypothesis, and the norm stabilization hypothesis. The norm broadcasting hypothesis is a mid-level functional hypothesis, sitting in between proximate and ultimate evolutionary hypotheses. The reputation signaling hypothesis and norm stabilization hypothesis are both at the evolutionary level, offered as two possible candidates for *why* the general norm broadcasting hypothesis might be the case.

We addressed each of the three hypotheses in one of the three studies of this chapter. In Study 4, we tested the prediction of the general norm broadcasting hypothesis that one's willingness to compensate was sensitive to whether one is being observed, as observation is necessary for the effectiveness of the signal. We found that compensators were in fact influenced by observers, consistent with the norm broadcasting hypothesis. This result mirrors that found in the domain of punishment, where the degree to which third parties were willing to sanction norm violators was shown to be sensitive to whether those third parties were observed (Kurzban et al., 2007).

We then moved on to address the two ultimate level functions for this behavior, the norm stabilization and reputation signaling hypotheses, both candidates to serve as the evolutionary support for the norm broadcasting hypothesis. We allowed for either, or both, of these hypotheses serve as the evolutionary rationale. In Study 5, we addressed the norm stabilization hypothesis: that by signaling the compensator's endorsement of the norm, compensation serves to stabilize the norm by mitigating the undermining signal of

the violation itself. This proposed rationale only functions if observers are in fact sensitive to witnessing compensation, increasing their propensity to conform. In support of the norm stabilization hypothesis, we found that participants who observed compensation were more likely to conform to the norm themselves. Additionally, we found that this effect was mediated by the observer's normative expectation that others thought they *should* conform to the norm, as predicted by the model. This finding is akin to that shown in the domain of punishment, where participants have been shown to be more likely to act pro-socially after observing punishment, even if they could not be subject to the same sanction (Fehr & Gachter, 2000; Stagnaro, Arechar, & Rand, 2017).

Lastly, in Study 6, we investigated predictions made under the reputation signaling hypothesis: that by signaling the compensator's knowledge and endorsement of the norm, compensation serves to better the compensator's reputation as an adherent to that norm, and therefore improve their attractiveness as a partner for future interactions. This proposal only functions if observers do in fact prefer compensators in future interactions. We failed to find support for the reputation signaling hypothesis, with observers selecting the compensator at a rate no different than chance. This pattern diverges from that observed in the punishment domain, in which people have shown a preference to interact with those who punish (Barclay, 2006; Jordan et al., 2016). Although this divergence could be due to distinct underlying mechanisms, it is important to first evaluate differences in study design. The experiments in the punishment domain showed that people prefer to interact with punishers *relative to non-punishers*. The parallel in compensation would plausibly be that people prefer to interact with compensators *relative to non-compensators*. This test is weaker than that which we evaluated in our study, in which we

compared a preference for compensators *relative to alternates who engaged in similarly pro-social acts other than compensation*. Previous studies allowed for the possibility that the preference for punishers over non-punishers was not driven by anything specific to the signal of punishment, but perhaps was driven by preferring someone willing to engage in any pro-social act. It is therefore possible that the reputational effects of compensation and punishment are similar, perhaps being taken as signals of general pro-social tendencies rather than fine-grained endorsements of the specific norm.

In summary, we found that people compensate more when being observed, and that observing compensation leads to higher degrees of conformity, but found no evidence for people preferring compensators as partners over similarly pro-social non-compensating alternatives. Taken together, these results are consistent with our hypothesis that compensation functions to signal the compensator's endorsement of the norm, and that this signal of endorsement functions as a norm-stabilization device. These empirical results parallel a number of similar findings in the punishment domain, suggesting that these two behavioral responses may be the result of one underlying cognitive mechanism responding to the violation of social norms. In this study, we found that our video manipulation of moral outrage increased compensation across situations, consistent with previous results suggesting anger to be a particularly difficult emotion to manipulate in isolation (Gross & Levenson, 1995). We found that the effect of the manipulation on compensation was mediated by a change in moral outrage *only* when the loss was due to a norm violation.

CONCLUSION

The compensation of victims can be found in groups ranging from hunter-gather to large-scale societies (Hill, et al., 2011; Rothaus, 2016). It is present at various social scales, from individuals compensating one another to enshrinement in various legal frameworks (Mullenix & Stewart, 2002; Palmer, 1979). It is observed both in and outside the lab (Charness et al., 2008). This research adds to our understanding of the psychological underpinnings of this phenomena. First, it provides a finer-grained understanding of the interaction between context and compensation, showing that moral outrage can motivate the compensation of norm violation victims, rather than empathic concern as previously argued. Second, it develops possible evolutionary accounts for this moral outrage driven compensation, and empirically tests the predictions of these accounts.

Previous work argued that victim compensation was motivated by empathic concern, whereas punishment was motivated by moral outrage (Coke et al., 1978; Fehr & Fischbacher, 2004). In Chapter 1, we showed that this understanding is incomplete, and provide a more nuanced picture. In Study 1 we demonstrated that, while trait disposition to feel empathic concern did correlate with a willingness to compensate the victims of bad investment decisions as demonstrated in previous work, the compensation of the victims of norm violations was uniquely predicted by a disposition to feel moral outrage. Additionally, we found that empathic concern and moral outrage were highly correlated, providing a plausible rationale for previous studies which demonstrated a relationship between empathic concern and the compensation of norm violation victims which failed to control for moral outrage.

Study 2 expanded the finding that moral outrage motivates compensation when the loss was due to the violation of a social norm beyond trait dispositions, into the domain of emotional states. In Study 2.a we directly manipulated moral outrage, and found that increasing moral outrage increased compensation across situations, consistent with previous work showing that anger is a particularly hard emotion to manipulate in isolation. We therefore tested the degree to which reported moral outrage *mediated* the effect of the manipulation on compensation, and found that moral outrage only mediated the effect on compensation when the loss was due to the violation of a social norm. We conducted a similar experiment in Study 2.b, but instead manipulated empathic concern. Here we found that increasing empathic concern did increase levels of compensation when the loss was due to the investment choice of the person experiencing the loss, but did *not increase* the compensation of the victim of a social norm violation. Taken together, these studies demonstrate on both the state and trait level that moral outrage is a unique driver of victim compensation, but only when compensating the loss of the victim of a norm violation.

We followed up this investigation with a replication of Study 2.a, in which we added monetary incentives as well as a modified dependent measure of compensation (Study 3). In this study, we did not find the previously observed effect of increasing moral outrage increasing the compensation of norm violation victims. Our best account of this finding relates to the particular dependent measure we adopted, which instead of measuring with what intensity the participant wants the victim to be restored, it measured the dollar amount the respondent thought would be correct. We therefore suspect that the null result observed may be the result of emotions such as moral outrage influencing the

intensity with which the victim wants the respondent restored, rather than what amount they deem to be fair restoration.

These results broaden our understanding of the emotional determinants of compensation, as well as the pro-social consequences of moral outrage. While previous work had identified empathic concern as the sole motivator for compensation (Batson et al., 1981; Coke et al., 1978; Toi & Batson, 1982), our studies reveal a richer landscape in which the particular emotional motivator is context specific, with moral outrage serving a critical role when a loss due to the violation of a social norm. These results also buttress the claim recently advanced in moral philosophy and psychology that despite anger often being described as an anti-social emotion (Averill, 1983), it often serves a pro-social purpose, such as deterring the intentional violation of norms (Gaus, 2011; Prinz, 2011; Russell & Giner-Sorolla, 2011). Here we see a new domain in which anger serves a prosocial good: driving the compensation of norm violation victims.

The findings described above led us to take up the question of the evolution of the compensation of norm violation victims. When two behaviors have distinct emotional determinants, they are taken to be the result of two distinctly evolved evolutionary systems. While others have theorized that compensation may act as a form of social insurance, sustained through indirect reciprocity, this unitary explanation does not account for our finding that compensation is driven by two different underlying systems, manifested in different contexts. While a general social insurance theory may account for empathic concern driven compensation, it fails to account for the moral outrage driven compensation we identified.

To address this shortcoming, we looked to the literature on the evolution of punishment. As punishment is driven by the same emotion (moral outrage) and occurs in the same context (a social norm violation), we found it plausible that the punishment of norm violators and the compensation of their victims may be driven by the same underlying process, both behavioral results of a single norm violation response system. We therefore expanded two models of punishment, showing how they might lead to the emergence of a norm violation response system which results in both compensation and punishment.

We first expanded on models which suggested that punishment could serve a costly signaling purpose, demonstrating the punisher's knowledge of local norms, a prerequisite for compliance. We argued that this same logic can be extended to compensation, where only one knowledgeable of local norms is capable of efficiently compensating what is actually a violation of a local norm. We also developed an alternative account, based on cultural group selection. Cultural group selection models rely on stable within group norms to reduce within group selection pressure, even when migration occurs between groups. Punishment maintains this within group stability by directly deterring violations, but also serves a pedagogical purpose, teaching those who observe the punishment what the normative beliefs of the group members are. While compensation cannot have the direct deterrent effect of punishment, it can serve a similar instructive purpose. These extensions provide a unique evolutionary accounting of the widespread behavior of compensating the victims of norm violations. Importantly, they also integrate two previously distinct areas of evolutionary theory, punishing victims and punishing their perpetrators.

To assess these accounts of the evolution of compensation, we derived novel predictions from each and tested them experimentally. Both proposed accounts rely on compensation to relay information (either the compensator's reputation or the norms of the community). In either case, these models therefore predict that a willingness to compensate would be conditional of the degree to which that compensation was observed. We tested this prediction in Study 4, observing that having participants make their compensation decisions publicly substantially increased participant's willingness to compensate. This effect is similar to that observed with punishment (Kurzban et al., 2007), consistent with a unified norm violation response system.

After testing a joint prediction of both models, we then separated them and developed a unique prediction to test from each. If compensation evolved to stabilize local norms within a group, then observing compensation would be predicted to increase compliance. To test this prediction in Study 5, we had participants choose whether to comply with a social norm, having just seen summary statistics of past participant's behavior. All statistics were held constant except for how many past participants chose to compensate a victim of the norm the participant would then choose whether or not to violate. We found that those shown high rates of compensation were more likely to comply with the norm than those shown low rates. This too is similar to effects observed for punishment, providing additional evidence for a unified norm violation response system (Fehr & Gachter, 2000; Stagnaro, Arechar, & Rand, 2017).

We then assessed the hypothesis that compensation functions as a signal of the compensator's status in the community and knowledge of the relevant local norms. This

hypothesis relied on observers preferentially interacting with compensators. We tested this prediction by allowing participants to choose from a set of possible partners, on whom they would have to rely on to conform to a reciprocity norm. One partner compensated a victim of the relevant reciprocity norm, one conformed to an unrelated norm, and one engaged in general pro-sociality. While we found that participants selected the generally pro-social partner less than by chance, it was the partner who conformed to an unrelated norm who was selected most, counter to our prediction. This result runs contrary to the underlying logic of our hypothesis that compensation signals specific knowledge of the relevant applicable norm, although it may indicate a new line of investigation, testing the degree to which observers believe conforming to a norm in one domain is predictive of conformist behavior in another.

Taken together, these results enrich our understanding of the emotional determinates and evolutionary roots of victim compensation. We expand the known motivators for compensation to include not only empathic concern but also moral outrage. Additionally, we demonstrate the interaction between these emotional drivers and social context, where moral outrage drives compensation only for victims of social norm violations. We built on existing models of compensation and punishment to account for these findings at an evolutionary level, suggesting an evolved norm violation response system, which includes both perpetrator punishment and victim compensation. Through experimental investigation, we then concluded that the strongest current evidence is for compensation to have emerged as a norm stabilization device, selected for through cultural group selection.

Our interpretation of the studies described above is not without limitations. The subject pools consisted entirely of WEIRD participants (Henrich, Heine, & Norenzayan, 2010). It is therefore only to that population that we can confidently generalize our results. To further substantiate the claim that humans evolved a particular cognitive mechanism, with particular emotional correlates, these findings must be conceptually replicated in a more diverse population. Of similar concern, we conducted our studies entirely in artificial settings, either in lab or via the internet, entirely within the trust game, both of which pose threats to external validity. To address this concern, we will need to expand our sphere of measurement less controlled environments to ensure that our results are not merely an artifact of our particular experimental setup.

The proposed norm violation response system provides an opportunity for a rich line of research to better understand the relationship uncovered between compensation and punishment. Although we propose that this system results in both punishment and compensation to serve a similar purpose, we have yet to investigate how these to behaviors interact, and in what context we might expect one versus the other. Past research has shown that punishment and compensation function as imperfect substitutes, where making one available reduces demand for the other (Chavez & Bicchieri, 2013; Jordan et al., 2016). Building on our understanding of the evolutionary costs of punishment and compensation, we may predict a number of relevant factors. As punishment can attract retribution (Chagnon, 1988; Cinyabuguma et al., 2006), we may predict that compensation will occur more frequently when the perpetrator is powerful and therefore can more easily counter-punish. Similarly, we may expect that while perpetrators may try to hide, victims may be much more readily available. We may

therefore predict that accessibility may account for a preference to compensate over punish in more realistic scenarios. Ecological observation may reveal additional factors driving both the choice to intervene after a norm violation, and if so, whether to help the victim or punish the perpetrator. Additionally, future work may integrate the social norm response system into formal models of norm propagation, providing empirical support and psychological realism.

REFERENCES

Amir, O., Rand, D., & Gal, Y. (2012). Economic Games on the Internet: The Effect of $1 Stakes. *PLoS ONE, 7*(2).

Averill, R. (1983). Studies on anger and aggression: Implications for theories of emotion. *American Psychologist, 30*(11), 1145-1160.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences, 111*(45), 15924-15927.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325-344.

Barclay, P. (2010). *Reputation and the Evolution of Generous Behavior.* New York: Nova Science Publishers.

Barclay, P. (Submitted). "Don't mess with the enforcer": Deterrence as an individual-level benefit for punishing free-riders.

Baron, J. (2007). *Thinking and Deciding.* Cambridge: Cambridge University Press.

Barr, A. (2001). *Social Dilemmas and Shame-based Sanctions: Experimental results from rural Zimbabwe.* Oxford: CSAE Working Paper No 2001-11.

Batson, C., Duncan, B. A., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology, 40*(2), 290-302.

Batson, D. (1991). *The Altruism Question: Toward A Social-Psychological Answer.* New York: Psychology Press.

Bekkers, R. (2006). Traditional and Health-Related Philanthropy: The Role of Resources and Personality. *Social Psychology Quaterly, 69*(4), 349-366.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior, 10*(1), 122-142.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

Bicchieri, C., & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making, 23*(2), 161-178.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making, 22*(2), 191-208.

Bowles, S. (2006). Group Competition, Reproductive Leveling, and the Evolution of Human Altruism. *Science, 314*(5809), 1569-1572.

Boyd, R., & Richerson, P. (1996). Why Culure is Common, but Cultural Evolution is Rare. *Proceedings of the British Academy, 88*, 77-93.

Boyd, R., & Richerson, P. (2009). Voting with your feet: Payoff biased migration and the evolution of group beneficial behavior. *Journal of Theoretical Biology, 257*(2), 331-339.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531-3535.

Boyd, R., Richerson, P., & Henrich, J. (2011). Rapid cultural adaptation can facilitate the evolution of large-scale cooperation. *Behavioral Ecology and Sociobiology, 65*(3), 431-444.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science, 6*(1), 3-5.

Carlsmith, K. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology, 42*(4), 437-451.

Carlsmith, K., Darley, J., & Robinson, P. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2), 284-299.

Chagnon, N. (1988). Life Histories, Blood Revenge, and Warfare in a Tribal Population. *Science, 239*(4843), 985-992.

Charness, G., Cobo-Reyes, R., & Jimenez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior and Organization, 68*(1), 18-28.

Charness, G., Naef, M., & Sontuoso, A. (2016). Self-Serving Conformism. *under review*.

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

Chudek, M., & Henrich, J. (2011). Culture-gene coevoltion, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences, 15*(5), 218-226.

Cialdini, R., Reno, R., & Kallgren, C. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015-1026.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics, 9*(3), 265-279.

Clutton-Brock, T., & Parker, G. (1995). Punishment in Animal Societies. *Nature, 373*(6511), 209-216.

Coke, J., Batson, D., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology, 36*(7), 752-766.

Cook, K., & Yamagishi, T. (2008). A Defense of Deception on Scientific Grounds. *Social Psychology Quarterly, 71*(3), 215-221.

Cronk, L. (2007). The influence of cultural framing on play in the trust game: a Maasai example. *Evolution and Human Behavior, 28*(5), 352-358.

Cushman, F. (2013). The Role of Learning in Punishment, Prosociality, and Human Uniqueness. In K. Sterelny, *Cooperation and Its Evolution* (pp. 333-372). Cambridge, Mass. : MIT Press.

Darwin, C. (1872). *The Expression of Emotions in Man and Animals.* London: John Murray.

Davis, M. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology,, 44*(1).

Davis, M., Mitchell, K., Hall, J., Lothert, J., Snapp, T., & Meyer, M. (1999). Empathy, Expectations, and Situational Preferences: Personality Influences on the Decision to Participate in Volunteer Helping Behaviors. *Journal of Personality, 67*(3), 469-503.

Dos Santos, M., Rankin, D., & Wedekind, C. (2011). The Evolution of Punishment Through Reputation. *Proceedings of the Royal Society B, 278*(1704), 371-377.

Drouvelis, M., & Grosskopf, B. (2016). The effects of induced emotions on pro-social behaviour. *Journal of Public Economics, 134*, 1-8.

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

Fehr, E., & Gachter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity. *The Journal of Economic Perspectives, 14*(3), 159-181.

Fehr, E., & Gachter, S. (2001). Altruistic punishment in humans. *Nature, 415*(6868), 137-140.

Fessler, D., & Haley, K. (2003). The Strategy of Affect: Emotions in Human Cooperation. In P. Hammerstein, *Genetic and Cultural Evolution of Cooperation* (pp. 7-36). Cambridge, Massachusetts: The MIT Press.

Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology, 30*(3), 263-276.

Fredrickson, B., & Levenson, R. (1998). Positive Emotions Speed Recovery from the Cardiovascular Sequelae of Negative Emotions. *Cognition and Emotion, 12*(2), 191-220.

Gaus, G. (2011). Retributive Justice and Social Cooperation. In M. White, *Retributivism Essays on Theory and Practice.* Oxford: Oxford University Press.

Gintis, H., Smith, E., & Bowles, S. (2001). Costly Signaling and Cooperation. *Journal of Theoretical Biology, 213*(1), 103-119.

Goerg, S., & Walkowitz, G. (2010). On the prevalence of framing effects across subject-pools in a two-person cooperation game. *Journal of Economic Psychology, 31*(6), 849-859.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology, 144*(4), 517-546.

Gross, J., & Levenson, R. (1995). Emotion elicitation using films. *Cognition and Emotion, 9*, 87-108.

Gross, J., & Levenson, R. (1997). Hiding feelings: The acute effects of inhibiting negative and positive emotion. *Journal of Abnormal Psychology, 106*(1), 95-103.

Gurven, M. (2004). To give and to give not: The behavioral ecology of human food transfers. *Behavioral and Brain Sciences, 27*(4), 543-559.

Guzman, R., Rodriguez-Sickert, C., & Rowthorn, R. (2007). When in Rome, do as the Romans do: the coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behaviot, 28*(2), 112-117.

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

Harris, J. (2006). Why the September 11th Victim Compensation Fund Proves the Case for a New Zealand-Style Comprehensive Social Insurance Plan in the United States. *Northwestern University Law Review, 100*(3), 1367-1407.

Hayes, A., & Preacher, K. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology, 67*(3), 451-470.

Henrich, J., & Boyd, R. (2001). Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas. *Journal of Theoretical Biology, 208*(1), 79-89.

Henrich, J., & Gil-White, F. (2001). The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural trasmission. *Evolution and Human Behavior, 22*(3), 165-196.

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*(5972), 1480-1484.

Henrich, J., Heine, S., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(29), 29.

Henrich, N., & Henrich, J. (2007). *Why humans cooperate: a cultural and evolutionary explanation.* New York: Oxford University Press.

Herrmann, B., Christian, T., & Gachter, S. (2008). Antisocial Punishment Across Societies. *Science, 319*(5868), 1262-1367.

Hertwig, R., & Ortmann, A. (2008). Deception in Experiments: Revisiting the Arguments in Its Defense. *Ethics and Behavior, 18*(1), 59-92.

Hill, K. R., Walker, R., Božičević, M., Eder, J., Headland, T., Hewlett, B., . . . Wood, B. (2011). Co-Residence Patterns in Hunter-Gatherer Societies Show Unique Human Social Structure. *Science, 331*(6022), 1286-1289.

Johnstone, R., & Bshary, R. (2004). Evolution of Spite Throgh Indirect Reciprocity. *Proceedings of the Royal Sociey B, 271*(1551), 1917-1922.

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

Jordan, K., McAuliffe, K., & Rand, D. (2017). Moral outrage by impartial observers: third party punishment is motivated by anger not envy or affective forecasting errors. *Under review*.

Kallgren, C., Reno, R., & Cialdini, R. (2000). A Focus Theory of Normative Conduct: When Norms Do and Do not Affect Behavior. *Personality and Social Psychology Bulletin, 26*(8), 1002-1012.

Kendal, J., Giraldeau, L., & Laland, K. (2009). The evolution of social learning rules: Payoff-biased and frequency-dependent biased transmission. *Journal of Theoretical Biology, 260*(2), 210-219.

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

Langergraber, K., Schubert, G., Rowney, C., Wrangham, R., Zommers, Z., & Vigilant, L. (2011). Genetic differentiation and the evolution of cooperation in chimpanzees and humans. *Proceedings of the Royal Society B, 278*(1717), 2546-2552.

Leliveld, M., van Dijk, E., & van 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.

Lerner, J., Goldberg, J., & Tetlock, P. (1998). Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin, 24*(6), 563-574.

Marshall, L. (1961). Sharing, Talking, and Giving: Relief of Social Tensions among !Kung Bushmen. *Africa, 31*(3), 231-249.

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

McElreath, R. (2003). Reputation and the Evolution of Conflict. *Journal of Theoretical Biology, 220*(3), 345-357.

Miller, G. (2000). *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature.* London: Anchor Books.

Mullenix, L., & Stewart, K. (2002). The September 11th Victim Compensation Fund: Fund Approaches to Resolving Mass Tort Litigation . *Connecticut Insurance Law Journal*, 123-137.

Nelissen, R. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242-248.

Nelissen, R., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions. *Judgment and Decision Making, 4*(7), 543-553.

Nettle, D., Panchanathan, K., Rai, T., & Fiske, A. (2011). The Evolution of Giving, Sharing, and Lotteries. *Current Anthropology, 52*(5), 747-756.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics, 92*(1), 91-112.

Nowak, M. (2006). Five Rules for Evolution of Cooperation. *Science, 314*(5805), 1560-1563.

Oliver, P. (1980). Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations. *American Journal of Sociology, 85*(6), 1356-1375.

Orne, T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*(11), 776-783.

Palmer, G. (1979). *Compensation for Incapacity: A Study of Law and Social Change in New Zealand and Australia.* Oxford: Oxford University Press.

Piazza, J., Russell, P., & Sousa, P. (2013). Moral emotions and the envisaging of mitigating circumstances for wrongdoing. *Cognition and Emotion, 27*(4), 707-722.

Pillutla, M., & Murnighan, K. (1996). Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers. *Organizational Behavior and Human Decision Processes, 68*(3), 208-224.

Poppe, M. (2005). The specificity of social dilemma situations. *Journal of Economic Psychology, 26*(3), 431-441.

Prinz, J. (2011). Against Empathy. *The Southern Journal of Philosophy, 49*(1), 214-233.

Raihani, N., & Bshary, R. (2015). The Reputation of Punishers. *Trends in Ecology and Evolution, 30*(2), 98-103.

Railhani, N., & Bshary, R. (2015). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution, 69*(4), 993-1003.

Rendell, L., Fogarty, L., Hoppitt, W., Morgan, T., Webster, M., & Laland, K. (2011). Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends in Cognitive Sciences, 15*(2), 68-79.

Rothaus, S. (2016, July 6). GoFundMe campaign raises $7 million for Orlando shooting victims. *Miami Herald*.

Russell, P., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion, 11*(2), 233-240.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379-399.

Schroeder, D., Steel, J., & Woodell, A. B. (2003). Justice Within Social Dilemmas. *Personality and Social Psychology Review, 7*(4), 374-387.

Skarlicki, D., & Folger, R. (1997). Retaliation in the workplace: The role of disributive, procedural, and interactional justice. *Journal of Applied Psychology, 82*(3), 434-443.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics, 87*(3), 355-374.

Stagnaro, M., Arechar, A., & Rand, D. (2017). From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition, 167*, 212-254.

Tennie, C., Call, J., & Tormasello, M. (2009). Ratcheting up the ratchet: on the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B, 364*(1528), 2405-2415.

The Washington Post. (2016, August 1). Minute by minute: How the attack in Orlando unfolded. *The Washington Post*.

Toi, M., & Batson, C. (1982). More evidence that empathy is a source of altruistic motivation. *Journal of Personality and Social Psychology, 43*(2), 281-292.

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

Tyler, T., & Boeckmann, R. (1997). Three Strikes and You Are Out, but Why? The Psychology of Public Support for Punishing Rule Breakers. *Law & Society Review, 31*(2), 237-266.

Wakslak, C., Jost, J., Tyler, T., & Chen, E. (2007). Moral Outrage Mediates the Dampening Effect of System Justification on Support for Redistributive Social Policies. *Psychological Science, 18*(3), 267-274.

Wendt, K., Frisina, L., & Rothgang, H. (2009). Healthcare System Types: A Conceptual Framework for Comparison. *Social and Policy Administration, 43*(1), 70-90.

Wenzel, M., & Thielmann, I. (2006). Why We Punish in the Name of Justice: Just Desert versus Value Restoration and the Role of Social Identity. *Social Justice Research, 19*(4), 450-470.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimonot, H., Horita, Y., . . . Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes, 120*(2), 260-271.

Zahavi, A. (1975). Mate selection—A selection for a handicap. *The Journal of Theoretical Biology, 53*(1), 205-214.