2012-09-05

The Roles of Empathy and Anger in the Regulation of Third-Party Punishment

Eric J. Pedersen
University of Miami, ericjohnpedersen@gmail.com

UNIVERSITY OF MIAMI

THE ROLES OF EMPATHY AND ANGER IN THE REGULATION OF THIRD-PARTY PUNISHMENT

By

Eric J. Pedersen

A THESIS

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science

Coral Gables, Florida

December 2012

UNIVERSITY OF MIAMI

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

THE ROLES OF EMPATHY AND ANGER IN THE REGULATION OF THIRD-PARTY PUNISHMENT

Eric J. Pedersen

Approved:

_____ Michael E. McCullough, Ph.D. Professor of Psychology	_____ M. Brian Blake, Ph.D. Dean of the Graduate School
_____ Debra L. Lieberman, Ph.D. Assistant Professor of Psychology	_____ J. Albert C. Uy, Ph.D. Associate Professor & Aresty Chair in Tropical Ecology

PEDERSEN, ERIC J. (M.S., Psychology)
The Roles of Empathy and Anger in the Regulation of Third-Party Punishment (December 2012)

Abstract of a thesis at the University of Miami.

Thesis supervised by Professor Michael E. McCullough.
No. of pages in text. (64)

Some researchers have recently promoted the idea that humans possess an instinct to "altruistically" punish social norm violators—that is, to intervene as unaffected third parties and punish transgressors at a personal cost, even when they have no hope of reaping any direct benefit from the punishment. Both the evolutionary theorizing and the empirical findings that are marshaled in support of this claim have been called into question, and results from a recent investigation strongly suggest that humans do not, in fact, punish altruistically as third parties on behalf of strangers. However, humans do engage in third-party punishment on behalf of people with whom they have a vested fitness interest, such as friends and kin. Herein it is proposed that empathy and anger are the proximate mechanisms that produce third-party punishment, and that they are only experienced when a third-party has a sufficiently high vested fitness interest in the victim of a transgression. Thus, the lack of third-party punishment on behalf of strangers can be explained proximately by the absence of empathy toward the victim and anger toward the transgressor. In a laboratory experiment with 212 participants, it was found that experimentally manipulated empathy felt toward a stranger increased anger at unfairness, and quasi-experimental comparisons with a previous experiment suggest that these differences in empathy and anger can produce third-party punishment. These findings provide preliminary support for the proposed model of third-party punishment and shed light on directions for future research on this topic.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

It is well established that humans have a penchant for punishing individuals who have harmed them directly, but some researchers have recently promoted the idea that humans also possess an instinct to "altruistically" punish social norm violators—that is, to intervene as unaffected third parties and punish transgressors at a personal cost, even when they have no hope of reaping any direct benefit from the punishment (Fehr & Fischbacher, 2004; Fehr & Gächter, 2002). Altruistic third-party punishment is a common feature of a body of theories that attempt to explain humans' "ultrasociality"—that is, their extensive cooperation with non-kin—by arguing that evolution has operated at both the individual and group levels to create a human psychology that is designed to motivate punishment of individuals who violate the cooperative norms of social groups. Here, this body of theory that is relevant to this proximate psychology will be referred to collectively as *group norm maintenance theories.* These theories propose that cooperative behaviors and punishment function to encourage cooperative gains globally for all the norm-abiding members of a group, while simultaneously making norm violation less profitable than norm upholding for all the members of the group. Though an individual may incur a personal cost by, for example, engaging in one-shot cooperation or third-party punishment, group norm maintenance theories argue that groups in which these behaviors are prevalent outcompete groups in which they are rare or absent, explaining (under certain conditions) the spread and maintenance of these behaviors (Fehr & Henrich, 2003; Gintis, 2000; Fehr, Fischbacher, & Gächter, 2002).

Though a growing body of experimental literature putatively supports the existence of third-party punishment in humans, the experimental economics paradigms

used in these studies contain design flaws that draw the validity of the results into question. Simply put, the widespread appearance of third-party punishment in such studies (e.g., Fehr & Fischbacher, 2004) is potentially due to experimental artifacts, including demand for punishment and audience effects (see 'Methodological features of the third-party punishment game' section). Using a modified version of the economic game typically used in third-party punishment research—one that removes several experimental artifacts—we have found that people are not willing to engage in third-party punishment, nor do they become angry at witnessing unfairness directed toward a stranger (Pedersen, Kurzban, McCullough, 2012). That is, people are unwilling to incur a cost to punish a transgressor if the transgressor has not harmed them directly: Of 66 research subjects that witnessed a player take money from another (the "victim"), only 2 (3%) invested enough money in punishment to reverse the selfish player's earnings at the victim's expense. Conversely, 15 (24%) of 62 subjects who were personally treated in the same manner retaliated enough to reverse the selfish players earnings.

On the basis of this eightfold difference in the rates at which bystanders and victims punish, then, it appears that people's willingness to invest their own resources into removing the gains that another individual has obtained by exploiting another person differs considerably as a function of whether the exploited individual is the self or a third party. Despite the elementary nature of this basic finding, we are not aware of any other research study that has investigated this question in humans. The goal of the present research is to investigate the emotional factors—namely, empathy and anger—that underlie punishment to proximately explain the differences between second- and third-party punishment.

Third-Party Punishment as "Revenge by Proxy"

Any theory that attempts to explain the existence of third-party punishment as a species-typical behavioral propensity of humans—that is the functional output of an evolved psychological mechanism—must account for how the associated fitness costs of punishment were outweighed by downstream fitness benefits in such a way that a species-typical propensity to punish could evolve. I propose that, rather than functioning as an altruistic benefit-delivery system, any mechanisms that produce third-party punishment are doing so as a form of "revenge by proxy" whose (ultimate) function is to acquire inclusive fitness benefits for the punisher by deterring aggressors from harming individuals with whom the punisher has a vested fitness interest. Consequently, third-party punishment should be deployed when the punisher stands to gain an indirect benefit that outweighs the cost of punishment. On this view, empathy for victims might be considered to be the output of an intermediate motivational system that computes the interdependence between the welfare of a victim and that of a prospective punisher—if welfare interdependence is sufficiently high, the system generates motivation to increase a victim's welfare by generating concern for victims and, as a result, anger and punishment toward harmdoers (Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008). That is, I view empathy and anger as proximate mechanisms for motivating third-party punishment, whose ultimate function is to acquire inclusive fitness benefits for the punisher. The present study represents the first step in testing this model by attempting to link empathy and anger with third-party punishment.

What Is Cooperation?

In the language of evolutionary biology, cooperation is defined as an act by an individual that benefits one or more recipients. Cooperative behaviors comprise two superordinate classes: Behaviors that are also beneficial to the actor (i.e., mutual benefit) and those that are costly to the actor (i.e., altruism). Importantly, benefits and costs apply to the lifetime direct fitness consequences of behaviors, not simply to a single interaction (West, Griffin, & Gardner, 2007). As West et al. (2007) suggest, it is useful to restrict the definition of cooperation to those behaviors that have evolved through natural selection specifically because of the benefit they provide to the recipient. For example, when a prey animal wanders toward a hidden predator it is benefiting the predator but, obviously, this behavior is not considered cooperative because the behavior is not produced by mechanisms that evolved to render benefits to predators.

More generally, classifying behaviors as cooperative requires not only benefit delivery *per se*, but it requires that the mechanisms generating those behaviors were designed by natural selection precisely because of the beneficial effects of their behavioral outputs on the direct fitness of others (and on the indirect fitness of the individual performing them). Indeed, when a prey animal detects a predator's location, it uncooperatively attempts to avoid being eaten—evidence that their inadvertent movement toward an undetected predator is not caused by mechanisms that were designed to deliver benefits to predators. Similarly, before labeling certain types of punishment in humans as altruistic (e.g., third-party punishment), one would want evidence not only that punishment is costly and delivers benefits to others, but that punishment is specifically designed for this purpose. That is, the observed costly delivery

of benefits is not merely a byproduct of other psychological mechanisms with functions other than the delivery of benefits to others (e.g., mechanisms for deterrence).

Cooperation: A Perennial Puzzle for Biology

Cooperation appears to be an evolutionary puzzle because individually costly cooperative behaviors often appear to reduce the fitness of the cooperator relative to non-cooperators. To the extent that this is actually the case, natural selection should work to undermine their evolution (West, El Mouden, & Gardner, 2011). However, given the tremendous prevalence of behaviors in which one individual behaves in ways that do, in fact, deliver benefits to others, it is reasonable to search for selection pressures that might, by boosting the benefit-provider's inclusive fitness (Hamilton, 1961), give rise to naturally selected mechanisms whose function is to deliver such benefits to others.

Solutions to the Puzzle. Broadly, cooperative behaviors can evolve if they provide indirect or direct fitness benefits to the actor (West et al., 2011).

Kin selection. Cooperative behaviors can confer indirect fitness benefits to the cooperator if the cooperative partner has a sufficiently high genetic relatedness to the cooperator (Hamilton, 1964). The process of kin selection explains how mechanisms that produce cooperative behaviors among genetic relatives can be favored by selection: Though the cooperator incurs a cost to provide a benefit to another, the relative has a high probability of possessing the genes that code for the cooperative behavior. The replication of these common genes can be promoted—thus, natural selection can favor mechanisms that produce behaviors that cause benefits for others—by increasing the reproduction of the partner, as long as the reproductive cost to the actor is less than the benefit to the partner, discounted by the probability that the partner shares those genes.

This condition can be stated mathematically with Hamilton's rule, $rb\text{-}c > 0$, where r is the coefficient of relatedness between two individuals (relative to the average relatedness of all individuals in a population), b is the fitness benefit delivered to the recipient of the behavior, and c is the fitness cost to the cooperator (Hamilton, 1963; 1964; West et al., 2011).

There are two main mechanisms by which r can be sufficiently high between individuals such that cooperative behaviors can be explained by kin selection: kin discrimination and limited dispersal (Hamilton, 1964; West, et al., 2011). Kin discrimination simply refers to the ability to distinguish kin from non-kin, whereby cooperative behaviors can then be directed towards kin (Hamilton, 1964). This discrimination can be based on environmental or genetic factors (Grafen, 1990) and has been demonstrated in many species, including humans (Lieberman, Tooby, & Cosmides, 2003; Lieberman, Tooby, & Cosmides 2007; West et al., 2011). Limited dispersal can also lead to a high relatedness among individuals, as relatives will be likely to remain in close proximity to each other (Hamilton, 1964). Thus, even if individuals cannot discriminate kin from non-kin, unconditional cooperation can still evolve in cases of limited dispersal, as group members are more likely to share a higher relatedness with each other than the population average (West et al., 2011).

Reciprocity. Direct fitness benefits can be accrued by cooperators through reciprocity, even when individuals are unrelated (Trivers, 1971). Whereas a single instance of incurring a cost to cooperate will lead to a relative fitness disadvantage for the cooperator, this cost can be overcome if it takes place in the context of a continuing cooperative relationship. This is nicely illustrated by the prisoner's dilemma: In a one-

shot interaction, defection always yields the highest average payoff—however, in the iterated prisoner's dilemma consisting of multiple rounds, mutual cooperation leads to higher average payoffs than mutual or alternating defection (Axelrod & Hamilton, 1981; Trivers, 1971).

Reciprocity is relatively unimportant for understanding cooperation among non-human animals (potentially from the presumably large cognitive demands involved in tracking social exchange; West et al., 2011, Clutton-Brock, 2009), but it seems to be very important in understanding human cooperation. For example, pooling risk and effort for hunting by sharing the spoils with others, on the condition that they reciprocate at another time, is much more efficient than hunting on one's own—solely for one's own kin—because hunting is a relatively high-variance method of food-acquisition. For example, among the Ache, a hunter-gatherer group currently living in eastern Paraguay, there is a 40% chance that a hunter will be unsuccessful in obtaining meat on any given day (Kaplan, Hill, & Hurtado, 1990). Thus, engaging in reciprocity can reduce or eliminate much of the variance associated with hunting and provide direct benefits to those involved (Cosmides & Tooby, 1989).

Though reciprocity can provide mutual direct benefits through gains in trade, it can expose the cooperator to exploitation (Cosmides & Tooby, 1989). Initiating a cooperative relationship with another individual by incurring a cost to provide them a benefit forms a social contract, but that contract may be violated by the other party since the exchange of services often does not take place simultaneously. Though exploitation is a possibility, reciprocity can still be favored by selection if there are mechanisms through which cooperative efforts can be directed toward likely cooperative partners

(e.g., kin detection; Lieberman, et al., 2007; Hamilton, 1964), away from cheaters (e.g., cheater detection; Cosmides, 1989; Tooby, Cosmides, & Price, 2006), or if cooperation can be enforced through punishment (West, et al., 2007).

What is Punishment?

Punishment is a response to the imposition of costs or the withholding of benefits that inflicts a fitness cost on the transgressor (Jensen & Tomasello, 2010; Clutton-Brock & Parker, 1995; McCullough, Kurzban, & Tabak, 2011). By definition, such an act is costly as it takes time, energy, and can put the punisher at risk of retaliation. Thus, for natural selection to favor the evolution of a mechanism whose function is to produce punishment behaviors, the behavioral outputs of such a mechanism must provide either direct (e.g., deterring future harm to one's self) or indirect (e.g., deterring future harm to one's kin) fitness benefits to the punisher.

Group Norm Maintenance and Altruistic Punishment

Based on several theoretical models and empirical studies, group norm maintenance researchers propose that some human punishment does not, in fact, provide direct or indirect fitness benefits to the punisher; rather, it is altruistic. Based on a model of the n-person prisoner's dilemma, Boyd and Richerson (1988) concluded that reciprocal cooperation could not evolve in groups larger than a few individuals because groups of cooperators would be vulnerable to invasion by noncooperators. However, when the possibility of punishing defectors was added to the model, cooperation could be sustained (Boyd & Richerson, 1992). Indeed, in this model, any behavior—including maladaptive ones—can be sustained with punishment if the cost imposed by punishment outweighs the cost of the behavior being enforced (Boyd & Richerson, 1992).

Gintis (2000) points to a similar pattern in empirical research utilizing the public goods game: Cooperation among people in a laboratory experimental economics game is significantly higher when punishment is allowed (Fehr & Gachter, 2000). Although punishment can lead to maintenance of cooperative behavior, it introduces a new problem: Whereas punishment of defectors in these examples benefits all group members equally, the cost of the punishment is borne only by those who punish. This creates the problem of second-order free riders who receive the benefits of punishment without bearing the cost (Henrich & Boyd, 2001; Fehr & Gächter, 2002; Boyd, Gintis, Bowles, & Richerson, 2003). Thus, selection could favor nonpunishers unless, in turn, they were punished for their second-order free riding. Clearly, this only adds another layer to the problem by leading to third-order free riders, and so on (Boyd & Richerson, 1992; Henrich & Boyd, 2001; West et al., 2011). Because of this infinite regress, group norm maintenance theorists assert that selection for cooperation and punishment must operate at a level higher than the individual such that the within-group fitness disadvantage that punishers face is compensated by selection between groups, as groups with a higher proportion of cooperators and punishers will have greater success than groups with a lower proportion. Indeed, several theoretical models have been proposed that demonstrate the plausibility of group norm maintenance (e.g., Gintis, 2000; Henrich & Boyd, 1998; Henrich & Boyd, 2001; Boyd, et al., 2003).

Economic Game Behavior as Putative Support for Group Norm Maintenance

It is well established that humans generally do not act in a so-called "rational" manner in experimental economics games; that is, they do not act in a way that maximizes their monetary gain from each interaction (Camerer & Thaler, 2007; Sally,

1995; Henrich et al., 2005). Deviation from perfectly selfish behavior in games in which an actor can impose costs upon another individual—but at a cost to the actor himself or herself—has been counted as evidence to support group norm maintenance theories[1]. Results from four types of game make up the bulk of the group norm maintenance literature: Dictator, Ultimatum, public goods, and third-party punishment.

Dictator Game. A clear example of a deviation from perfect selfishness is the typical behavior of players in the Dictator Game. The Dictator Game is a simple interaction that consists of two anonymous participants: one is assigned the role of the so-called Dictator, the other the role of the so-called Receiver. The Dictator is given a sum of money and is instructed to divide it between the two subjects however he or she pleases, while the other player merely receives any funds the Dictator transfers and has no influence on the interaction. Because the game takes place anonymously and players are instructed that there is only one round—thus, retaliation is not possible—the rational strategy for the Dictator is to transfer none of the money to the other player, thereby maximizing the Dictator's monetary payoff. This strategy, however, is rarely observed in actual experiments: A recent meta-analysis revealed that Dictators typically choose to transfer about 28% of their stake (Engel, in press).

---

[1] The use of games involving money can obviously only serve as a proxy for understanding the evolution of psychological mechanisms for punishment. Though it may be unrealistic to equate the loss of a few dollars in a laboratory experiment to losing one's reputation in a group, behavior in such games can still be used to test hypotheses regarding the proximate psychological mechanisms responsible for producing punishment, based on theoretical predictions of their ultimate function. That is, despite the [presumably large] difference in magnitude in losing dollars versus reputations, relative differences (e.g., losing dollars versus no losing dollars) can still shed light on the underlying psychological processes. Furthermore, there is evidence that behavior in economic games does not vary much based on the size of the stakes involved (Cameron, 1999; Carpernter, Verhooden, & Burks, 2005), suggesting that, even when the stakes are small, the underlying psychological processes are very similar to much more serious situations.

Ultimatum Game. The Ultimatum Game (Güth, Schmittberger, & Schwarze, 1982) is a modified Dictator Game in which the second player is allowed to reject the split the first player proposes. Because rejection causes each player to earn nothing, it has been interpreted as a form of "costly punishment." Rejections in the Ultimatum Game have been interpreted by group norm maintenance theorists as altruistic because the responder incurs a personal cost to punish the proposer for a violation of a social norm and, as a result, the proposer may adjust his or her behavior to conform to the norm in the future by treating members of the social group more fairly—thus, the punisher indirectly benefits others at a personal cost (Fehr and Fischbacher, 2003).

Clearly, rejection does not maximize one's monetary gain on a given interaction, but results from the Ultimatum Game shed light on the role punishment can play in promoting cooperation: Proposers in the Ultimatum Game typically offer the second player around 40% of their stake as opposed to the 28% that is typical in the Dictator Game (in which punishment is not possible). Although there is some between-culture variation in the average amount transferred and rejected in these games, cross-cultural research has found no culture in which the completely rational strategy of money maximization on individual trials is commonly employed (Henrich et al., 2005; Henrich et al., 2006; Oosterbeek, Sloof, & van de Kuilen, 2004). Furthermore, every culture tested has shown a willingness to impose costly punishment in the Ultimatum Game (Henrich et al., 2006): In each of the 15 world societies in which Henrich et al. examined ultimatum game behavior, second players' likelihood of rejecting proposers' offers increased as the proportion of the stake offered to the second players decreased from 50%.

Third-party punishment game. Similar results have been found in the third-party punishment game, which is a modified version of the Dictator Game (Fehr & Fischbacher, 2004; Henrich et al., 2006). The game consists of a "Dictator" who chooses to give any portion of a sum of money (or nothing at all) to a "Receiver," who has no influence on the interaction. A third player is assigned the role of "Adjuster" and instructed that they may adjust (for a cost) the outcomes of the two other individuals following their economic interaction. Other than having the ability to adjust the outcome, the Adjuster is uninvolved in the interaction; regardless of the Dictator's decision, the only way the Adjuster's monetary outcome can be affected is if he or she decides to punish the Dictator. With the game structured in this manner, people appear to punish altruistically as third parties: Adjusters typically incur a personal cost to punish Dictators for unfair splits of the money, regardless of the fact that the Adjusters cannot derive financial benefit from punishment, nor are they personally harmed by the transgression. Punishment in the third-party punishment game is critically different than in the Ultimatum Game—in which a punisher is directly affected by an unfair split of money by the proposer—and in the public goods game (see below)—in which low contributions from free riders lower the payoffs of all other members of the group. Thus, people who punish in the Ultimatum Game and public goods game have done so after they personally have been affected; those who punish in the third-party punishment game have only witnessed unfairness directed towards someone else, suggesting that people are willing to intervene—at a cost—on behalf of others, even if they personally have not been harmed.

Public goods game. When experimental economics games take place in a group setting, such as in a public goods game (Olson, 1965; Yamagishi, 1986), behavior also

appears to not maximize selfish interests (Fehr & Gächter, 2000). In the public goods game, each player in a group is given an endowment by the experimenter, which they are free to divide in any way between their private account and a group account. Any amount donated to the group account is multiplied by the experimenter, and then divided equally between all members of the group without regard to those group members' donations. When given the opportunity to punish other group members at a cost to oneself (for example, by including a punishment round in which group members may pay a small cost in order to "burn" money from other group members' accounts), many players do so when they perceive that other players contributed an unfairly small amount. Because any benefit produced by punishment in such a design—such as increasing future cooperation from those who contributed little—benefits everyone in the group, punishers and non-punishers benefit equally from an act of punishment, but only the punisher incurs a cost. This means that the payoff for punishing defections is lower than the payoff for not punishing them; that is, punishers incur a cost relative to non-punishers. In games allowing punishment, there are higher levels of cooperation within the group, even though the average individual payoff within the group is smaller for punishers relative to non-punishers (Fehr & Gächter, 2000).

In a widely cited 2002 paper, Fehr and Gächter ran a series of public goods games in which subjects played twelve rounds of one-shot games with completely orthogonal sets of players: They were assured that they would never encounter the same player twice. In one condition, costly punishment was allowed at the end of each of the first six rounds, followed by six rounds without punishment. In the other condition, the first six rounds were played without punishment and followed by six rounds that included

punishment. Subjects frequently punished when the option was available and cooperation in those rounds was higher than in rounds without punishment, regardless of the order of the rounds. Since subjects would never again encounter the players they had punished, thereby eliminating any chance of direct benefits, Fehr and Gächter considered punishment in this paradigm altruistic; they proposed that punishment was employed against low contributors to prevent them from behaving in the same way in the future, thus benefiting the players that subsequently interact with them (Fehr & Gächter, 2002).

Deconstructing the Evidence Supporting Group Norm Maintenance

Though the literature supporting group norm maintenance has proliferated over the past decade, several widespread theoretical and empirical issues call into question the conclusions drawn from this research.

Problems with theoretical models of altruistic punishment. A common theme throughout the group norm maintenance literature is to state that standard explanations for the evolution of cooperation—namely, reciprocity and kin selection—are necessary but not sufficient to explain the level of cooperation found in humans; rather, some form of group selection is also needed (e.g., Gintis, Bowles, Boyd, & Fehr, 2003). Several group norm maintenance models have been proposed to support the assertion that group selection is necessary to explain human cooperation but certain aspects of the models call this claim into question.

Oversimplification. Arguably, much weight has been attributed to oversimplified theoretical models that seem to have actually created some of the problems that theories of group norm maintenance attempt to solve. For instance, Boyd and Richerson's (1988) model using the n-person prisoner's dilemma is cited throughout the literature as

evidence that reciprocity likely cannot explain cooperation in large (i.e., more than a handful of individuals) groups (e.g., Gintis, 2000; Gintis, et al., 2003; Fehr & Fischbacher, 2003; Henrich & Boyd, 2001; Henrich et al., 2005; Richerson & Boyd, 1997; Boyd, et al., 2003; Boyd, Gintis, & Bowles, 2010). However, this model is limited in its real-world applicability because (a) it does not account for relatedness among group members; (b) it assumes that all interactions take place globally (i.e., dyadic or small-group interactions within the larger group are not possible); and (c) it assumes that the only form of retaliation an individual can take against free riders is the withdrawal of cooperation—an act that does not specifically target a cheater but harms the entire group instead. Boyd and Richerson (1988) noted (c) as a limitation, and then addressed it with a follow-up model that allowed for punishment (as distinct from cooperation) targeted at specific individuals (Boyd & Richerson, 1992). With this model, they found that punishment targeted specifically at a cheater can indeed sustain cooperation and, importantly, that costly punishment can even provide direct benefits to the punisher if the punishment reforms the free rider (Boyd & Richerson, 1992). However, they still suggested that cooperation likely cannot evolve in large groups without the aid of altruistic punishment—and therefore, that some form of multilevel (i.e., between-group) selection is required. As a result, this assertion is often used as a basic assumption in more recent models (e.g., Gintis, 2000; Boyd et al., 2003).

Inclusive fitness. Inclusive fitness is often either misunderstood or overlooked in group norm maintenance models. For example, models of group norm maintenance are proposed as an explanation for the vast amount of cooperation in humans among unrelated individuals that putatively could not have evolved as a result of kin selection

(Gintis, 2000; Bowles & Gintis, 2004). However, these models assume that groups are both relatively small and migration is infrequent—precisely the conditions that lead to high relatedness among individuals via limited dispersal—thus eliminating the need to invoke group selection (Hamilton, 1964; West, et al., 2011). Indeed, no model of group norm maintenance—or any other formal model relying on group selection—has been proposed that cannot also be explained in terms of kin selection (West, et al., 2011).

Additionally, theories and models proposing altruistic punishment often overlook the direct and indirect benefits that can accrue to punishers that enable them to recoup the costs of punishing, thus eliminating the need to explain the evolution of punishment at a level higher than the individual. Indeed, all known examples of punishment in the non-human animal literature can be explained in terms of direct and indirect benefits (Clutton-Brock & Parker, 1995; Smith et al., 2010). Models of cooperation in humans have shown that punishment can lead to direct and indirect benefits to the punisher, provided that cooperation is facultative—that is, individuals will adjust their level of cooperation based on the threat of punishment or in response to being punished (Boyd & Richerson, 1992; Gardner & West, 2004).

Lack of evidence for altruistic third-party punishment in the real world. In the non-human animal literature, evidence clearly suggests that third-party punishment is not altruistic: A review of 98 published studies of 49 non-human species revealed that third parties regularly become involved in other individuals' conflicts, and that these third-party interventions can be explained, in all cases, in terms of the inclusive fitness benefits to third parties through helping their kin, helping individuals from whom they

can expect reciprocal help in the future, or obtaining direct fitness benefits (Smith, et al., 2010).

Outside of experiments using economic games, evidence for altruistic third-party punishment among humans is also lacking. Indeed, as is the case with the literature on non-human animals, the evidence clearly suggests the vectors by which third-party punishers might accrue inclusive fitness benefits. For instance, a study of 200 violent and matched non-violent conflicts reported by 100 male prisoners revealed that in only 1% of the reported conflicts did a third party without friendship, family, or gang ties violently intervene in the conflict. However, the odds of third-party intervention were 14 times higher if the third party was a friend of one of the disputants and 32 times higher if the third party was a family member or fellow gang member of one of the disputants (Phillips, Cooney, Carr, & Frady, 2005).

Similarly, Lieberman & Linke (2007) found in a hypothetical vignette experiment that subjects recommended longer sentences for burglars who targeted subjects' family members than those who stole from schoolmates or foreigners. Furthermore, in a second study, subjects reported that they personally would be willing to sacrifice 13 days without pay to help find a burglar who targeted a family member whereas they would only sacrifice 2 days on behalf of a schoolmate or foreigner. Taken together, this evidence suggests that without the availability of the sorts of direct benefits that might come from helping an alliance partner, or the indirect benefits that might come from aiding a genetic

relative or family member, costly third-party punishment in the real world—at least in violent settings—is rare[2]

Problems with altruistic punishment interpretations of economic games. Though it has been shown that punishment can directly benefit the punisher and that evidence in support of altruistic third-party punishment is lacking in other contexts, the fact that people do indeed appear willing to impose punishment on non-cooperators in experimental economics games—even when imposing such punishments requires them to pay a cost to do so—needs to be explained. In experimental games, cooperation often deteriorates without punishment but can be sustained when punishment is allowed. Within the context of these games, subjects who punish appear altruistic to strangers as they incur personal costs they cannot overcome to benefit other group members. But is this really altruism?

'Altruism' as a design artifact. A large source of confusion in the cooperation literature, both in human and non-human animals alike, can be attributed to the misuse of the term *altruism*, including Trivers' (1971) initial conceptualization of reciprocity, which he termed 'reciprocal altruism' (West, et al., 2007; West, et al., 2011). Technically, altruism is defined as a behavior that is costly to the actor and beneficial to one or more recipients, *with costs and benefits defined as the lifetime direct fitness consequences of the action* (West, et al., 2007). Thus, reciprocal altruism is not

---

[2] Third-party punishment on behalf of [pair-bonded] mates should also be common, as mates can provide both direct and indirect benefits. To the best of my knowledge, no third-party punishment research has addressed this possibility as of yet.

technically altruism at all, as costs are more than repaid through subsequent interactions as both individuals mutually benefit from cooperation.

The consideration of lifetime fitness consequences of behaviors is important in the context of experimental games as the games are stylized to lend them themselves to a laboratory setting. Much has been made of the irrationality of subjects' typical decisions—and terming them *altruistic*—in anonymous, one-shot interactions in such games which may, in fact, be a misleading byproduct of well-designed systems for cooperation. Anonymous, one-shot interactions were likely either rare or non-existent over the course of human evolutionary history and should not have been a strong selective pressure on shaping behavior in social interactions (Hagen & Hammerstein, 2006). Thus, behavior in social interactions might have been shaped by natural selection for engaging in repeated interactions, usually with either kin or relatively well-known individuals. In the interest of experimental control, subjects in laboratory experiments are often exposed to interactions that, presumably, do not mimic the contexts for which mechanisms that produce punishment behavior—subjects' actions in experimental games, then, may not be indicative of the selected function of the behavior-generating mechanisms in question.

For example, punishment appears altruistic in Fehr & Gächter's (2002) public goods experiment because only one round is played with any given set of individuals before the groups are changed and therefore can only benefit future interactants with the punished individual. However, the putative altruistic punishment in this case could plausibly be due to the strange design of the game: The rotation of groups clearly has no real-world analogue and prevents even the *possibility* of a punisher personally gaining

benefits from punishing (i.e., higher payoffs on subsequent rounds based on a reformed free rider's contributions). Furthermore, punishment takes place on the same round—with the same group of people—that transgressions occur, so punishment is directed against those who have affected the punisher's personal earnings on the round; that is, punishment in this case is very clearly retaliatory, regardless of whatever beneficial effects in may have on the transgressors future interactants. Therefore, such behavior could be the result of mechanisms whose function is to deter individuals who have harmed the punisher directly rather than mechanisms designed to deliver benefits to others.

A much stronger case for altruistic punishment could be made with a public goods game in which the punisher is not personally affected by the outcomes. Carpenter and Matthews (2005) ran a public goods game that featured such a condition and found only minimal support. In a one-shot public goods game, subjects could punish people within their own group and in other people's groups but could not be punished by those in other groups. In this condition, only 10% of subjects engaged in third-party punishment (i.e., punishing outside of their own group) and the average amount invested such punishment was $.10. However, the same subjects invested approximately seven times as much money to punish low contributors within their own group.

Methodological features of the third-party punishment game. Another cause for the apparent altruistic punishment in experimental games could be methodological in nature. Here I will focus specifically on the third-party punishment game, though some of these criticisms are applicable to other games as well, including the public goods game. Two major issues with the standard design of the third-party punishment game may

create the appearance of a human motivation to altruistically punish as third parties, even if none actually existed: experimental demand characteristics and audience effects.

Experimental demand characteristics. The third-party punishment game is structured in a way that may artificially encourage third-party punishment as participants in the Adjuster role are only given one option to actively participate in the game: third-party punishment. Adjusters are often prompted by a question asking them how much they would like to adjust the outcome and, although they can choose not to punish at all, the very nature of the question implies that the participant's task is to—at least some of the time—adjust the other individual's outcome. Participants who assume the role of the "good subject" (Weber & Cook, 1972), therefore, can be expected to search for a rule that would enable them to vary their behavior in accordance with some varying aspect of the experimental situation; for example, how much the putative target of their third-party punishment has failed to share with another individual.

Furthermore, the third-party punishment game is frequently played using what is called the strategy method, which consists of the Adjuster choosing how he or she would respond to each possible choice of the Dictator before the actual choice is revealed (e.g., Fehr and Gächter, 2004). This method is used to enhance statistical power because economics experiments are typically conducted without the use of deception. Consequently, there will inevitably be combinations of actions that occur infrequently and the strategy method allows for analysis of such situations. However, this method involves experimental artifacts that may influence results of the third-party punishment game in undesirable and unintended ways. First, since decisions to punish are made *before* the Dictator's decision is revealed, emotional—or other—reactions to the

Dictator's action are completely eliminated from the decision to punish. Second, the strategy method may produce inflated levels of punishment as players must make a decision to punish or not for each possible choice of the Dictator instead of simply responding to one action. Thus, they are asked repeatedly how they would respond to different actions—some of which are clearly more unfair than others—which could lead subjects to produce biased answers based on a similar principle as the anchoring and adjustment heuristic (Tversky & Kahneman, 1974), wherein decisions to punish different splits of the stake are affected by the initial scenario with which they are presented. This is an especially nettlesome design flaw in an experiment in which *any* amount of punishment is deemed supportive of the altruistic punishment hypothesis.

Audience effects. Inherent to the design of the third-party punishment game is the constant presence of a witness to any interaction that occurs between any two individuals. The presence of such an audience can affect one's willingness to punish as it introduces variables other than the direct costs and benefits of punishing—namely, reputational considerations that could accrue indirect fitness benefits. When reputational gains are taken into account, punishment directed at a wrongdoer who harmed a third party could function as a type of costly signal. There are several possibilities for what information a willingness to punish third parties can signal, including quality as a mate (Zahavi, 1975; Kurzban, Descioli, & Obrien, 2007), quality as a cooperative partner (Fessler & Haley, 2003), and formidability to prevent future exploitation of oneself (Johnstone & Bshary, 2004) or one's friends and kin (Lieberman & Linke, 2007). Thus, third-party punishment observed in the game may be motivated by *strategic attempts to signal one's willingness to punish violations as a means of indicating how one would behave if personally*

harmed, rather than by *one's desire to enforce norms*. Indeed, it has been shown—though with a different paradigm—that observers of unfair treatment punish third parties significantly less when they are assured no one will see their decision (Kurzban et al., 2007).

Results from a Modified Third-Party Punishment Game

Our laboratory designed a modified version of the third-party punishment game (described in greater detail in the Methods section below; Pedersen et al., 2012) that removes some of the aforementioned methodological problems of previous research on third-party punishment. In a study of 275 participants with this new paradigm, people were reticent to engage in third-party punishment; they were much more likely to retaliate when they were personally mistreated. Twenty two (35%) of 62 victims of unfair behavior (those from whom a harmdoer took $4 of a $5 endowment) punished on their own behalf—15 of whom punished at least $4, which was sufficient to reverse the harmdoer's unfair gains. In contrast, only 11 (17%) of 66 witnesses of unfair behavior (those that observed a harmdoer take $4 of a $5 endowment from another) punished on behalf of the victim, and, of those that punished, *only 2* punished at least $4. On average, receivers of unfairness incurred a cost of around $.30 to punish the transgressor (of a possible $1.25; punishment cost subjects .25 times the amount punished), whereas witnesses of unfairness only incurred a cost of around $.10. I propose that this difference in punishment can be explained proximately by a difference in anger elicited by the two cases—specifically, that witnesses of unfairness are substantially less angry at the transgressor than receivers of unfairness.

The Role of Anger in Punishment

Anger has been strongly implicated in human punishment (Fehr & Gächter, 2002; Petersen, Sell, Tooby, & Cosmides, 2010; Nelissen & Zeelenberg, 2009; Srivastava & Espinoza, 2009; Jensen, 2010) and thus likely plays a large role in the maintenance of cooperation. Sell's (Sell, Tooby, & Cosmides, 2009; Sell, 2011) recalibrational theory of anger proposes that anger's function is to motivate bargaining tactics (including punishment) that resolve conflicts of interest between parties when one party has displayed a lower valuation of the other's welfare (e.g., unfairly splitting a resource) than is deemed appropriate. This type of bargaining system can help to maintain cooperation through punishment—or threat of punishment—by preventing future exploitation and recalibrating a transgressor's valuation of the punisher so that mutually beneficial interactions can take place in the future. Following the logic of the recalibrational theory, I propose that this view of anger provides a possible explanation for why people may fail to punish on behalf of unrelated anonymous strangers: The uninvolved third party has not been subjected to a conflict of interest (i.e., the third party has not picked up on a cue that was ancestrally associated with a negative impact on inclusive fitness).

Evidence for the role of anger in third-party punishment. Indeed, research has shown that anger at unfairness is only strongly evoked when the target of unfairness is the self or a close other; unfairness to strangers does not evoke much anger, even though appraisals of the morality and fairness of the situations are equal (Batson et al., 2007). Results from our previous study with the modified third-party punishment game (Pedersen et al., 2012) replicate this finding: Self-reported judgments of both the fairness and moral-wrongness of unfair treatment were identical regardless of the target of

unfairness in the interaction (consistent with other third-party punishment research; Lieberman & Linke, 2007). However, subjects' self-reported anger toward the offending party was significantly higher when they personally experienced unfairness than when they merely observed the unfair treatment of another individual. Results from an implicit measure of anger (a lexical decision task, described in the methods section) also support this conclusion: Participants who received unfair treatment were quicker to correctly identify hostile words (e.g., anger, kill) than were witnesses of unfairness—who were no quicker to identify these words than subjects in control conditions—suggesting they were indeed angrier (Gollwitzer & Denzler, 2009). Thus, the marked difference between third-party and second-party punishment seems to be due to a lack of anger toward the provoker rather than to differences in attention, moral judgment, or conceptions of fairness of the situation.

In the previous study, observers of unfairness reported (via self-report) a small amount of anger that was significantly greater than observers of fairness. This may shed light on the small amount of third-party punishment found in the preliminary study: When self-reported envy and jealousy were controlled for, observers of unfairness reported no more angry than observers of fairness, whereas receivers of unfairness remained significantly more angry than receivers of fairness. Thus, it appears that the anger reported by witnesses of unfairness in this study was attributable to envy of the transgressor's unfairly-gained extra money—before any punishment takes place, the subject in this situation will have $5 compared with the transgressor's $9—and not to the unfair behavior per se.

The Present Study: Does Empathy Elicit Anger and Third-Party Punishment?

Third-party punishment and anger at witnessing unfair treatment are either minimal or nonexistent when the victim is a stranger, which is in stark contrast to when the self is the target of unfairness. Following the logic of Sell et al.'s (2009) recalibrational theory, I propose that non-envious anger is only evoked when the inclusive fitness of an individual is infringed upon—either in the self or an individual in which the self has a vested inclusive fitness interest; the lack of anger at witnessed unfairness, relative to experienced unfairness, may be due to a low amount of welfare interdependence with the victim (i.e., the degree to which the fitness outcomes of a victim impact the inclusive fitness of the witness). Thus, anger and third-party punishment may be dependent on the experience of empathy toward the target of unfairness, as empathic anger is generally only invoked when one feels close (either due to kinship, friendship, or shared interests) to the harmed party (Batson et al., 2007; O'Mara, Jackson, Batson, & Gaertner, 2011). For this view of third-party punishment to be tested, it first needs to be established that empathy plays a causal role in the elicitation of anger and third-party punishment[3]. Can manipulating the empathy third parties feel toward targets of unfairness lead to an increase in anger, thereby eliciting substantial levels of third-party punishment?

Some experimental evidence supports this idea. In addition to finding that witnessing unfairness directed towards a stranger did not elicit anger, Batson, et al.

---

[3] Note that this view of third-party punishment requires the establishment of empathy's proximate role in producing anger and third-party punishment; should this role not be supported, the hypothesis cannot be correct. Though my predictions are derived from principles of inclusive fitness and kin selection, the present experiment cannot directly speak to their roles in the ultimate function of third-party punishment.

(2007) found that experimentally-induced empathy towards a stranger could elicit self-reported anger in third parties after observing the stranger be mistreated. Similarly, Vitaglione and Barnett (2003) found that a self-report measure of trait empathic anger—a measure of one's proneness to feel anger on behalf of mistreated others—was associated with subjects' self-reported likelihood of investing effort to punish a drunk driver after the subjects listened to a (sham) recording of the victim of the accident the driver caused. Thus, empathy for victims seems to play an important role in eliciting anger and possibly motivating third-party punishment, and experimentally induced empathy can have these effects, even when the targets of unfairness are anonymous.

For the purposes of this project, I propose that empathy is an output of a mechanism that takes welfare interdependence into account and is elicited when one witnesses unfair treatment directed towards a person with whom the witness has a vested inclusive fitness interest. On this view, empathy should not be elicited when anonymous strangers—with whom third parties share little, if any, welfare interdependence—are harmed. As welfare interdependence (i.e., one's inclusive fitness interests) with the victim increase, empathy should increase, thereby eliciting anger and motivating third-party punishment. Because of this, experimentally induced empathy may act as an artificial cue that one's welfare interdependence with the victim of a transgression is such that warrants retaliatory action because one's own inclusive fitness has been harmed by the transgression. That is, people may retaliate on behalf of strangers for whom they feel empathy, which is probably a byproduct of the function empathy (presumably) evolved to serve.

The present study used a well-established experimental manipulation of empathy (Batson et al., 2007) to investigate whether increases in empathy toward a victim of unfairness can decrease or eliminate the differences in both self-reported and implicit anger in cases of second- and third-party transgressions, and reduce or eliminate the difference in third- and second-party punishment.

Chapter 2: Method

Participants

Participants were 212 (116 female) University of Miami undergraduates (M_{age} = 18.80, *SD* = 1.87) enrolled in introductory psychology courses. They were recruited through the psychology department's research participation pool and participated for partial course credit and $9 in compensation. Sessions lasted one hour.

Design

This between-subjects design had three groups that witnessed unfairness with an empathy prime (Low-Empathy, High-Empathy, No-Instructions) and an offset control group that received unfairness (No-Instructions). Participants were randomly assigned to one of the four conditions.

Procedure

After obtaining consent, the experimenter initiated a computer program that ran the experiment in E-prime (version 2.0). The program provided all of the instructions for the game and tasks and guided the participant through the study. Participants were told they were interacting with two other players over the computer network in an economic decision-making game that would last for multiple rounds and that they would be paid based on the money they earned during the game. In reality, they interacted with a preprogrammed computer script and were paid a fixed sum. After the rules and roles of the game were explained to the participant, they were also informed that their session would either involve communication or no communication. In a communication session, one participant would be randomly assigned to be the sender of communication and one participant would be randomly assigned to be the receiver of communication. In a no

communication session, no one would write or read a note. All participants were (ostensibly randomly) selected to be receivers of communication. Participants in the low-empathy and high-empathy conditions were then instructed to read the sender's paragraph while taking a certain perspective.

Empathy manipulation. Participants in the low-empathy condition were asked to take an objective perspective towards what was described in the note. Participants in the high-empathy condition were asked to try and imagine how the sender of the note felt about what was described. The note itself described the sender as being saddened after a recent break-up with a significant other (Appendix A). Both the letter and the method have been successfully used as an empathy manipulation in many experiments (e.g., Batson, et al., 2007). As a manipulation check, participants were probed for their emotional reactions (e.g., empathic, sympathetic) to each of the players following each round in the game.

Decision-making game. Following the empathy manipulation, the economic game began. The game consisted of two rounds in which each player was given $5 to use in each round and assigned to one of three roles: The *Decision-Maker* who ostensibly had the option to give any portion of her $5 to the *Receiver* or take any portion of the Receiver's $5; the *Observer* merely saw the results of the round and was unaffected by the Decision Maker's choice. Participants were randomly assigned to be either the Observer or the Receiver in the first round and the (computer-programmed) Decision-Maker chose to take $4 from the Receiver. The computer then showed a summary screen for the round that detailed the amount of money each player earned for the round.

Following the round, participants completed a lexical decision task (detailed below) and a series of self-report questions (detailed below).

Prior to role assignment for the second round, participants were informed that there would be no Observer in Round 2; one player would be assigned to a different task and be unable to see the results of the interaction. All players were given another $5 and the participant was chosen for the role of Decision-Maker while the Decision-Maker from Round 1 was assigned the role of the Receiver (ostensibly by chance). Participants were instructed that they could give any amount of money to the Receiver, do nothing, or remove money from the Receiver's account (the word "punishment" was not be used). Removing money cost one-fourth of the amount removed and, unlike in the first round, would not be gained by the participant; it disappeared (making this a costly punishment choice). A second lexical decision task followed, along with the same self-report questions as in the first round. Participants were then debriefed through an extensive, staged process (on the computer) to assess the credulity of the experiment and to explain why deception was necessary (Aronson, Ellsworth, Carlsmith, & Gonzales, 1990).

Measures

Choice as decision-maker in second round of the game. Participants were allowed to transfer any amount of their stake to the Receiver, deduct any amount from the Receiver's $5, or do nothing. Deducting money cost the participant 25% of the amount to be deducted from the Receiver, and money deducted was not transferred to the participant—it simply disappeared. Thus, deducting money served as a measure of costly punishment with no material gain to the punisher.

Lexical decision task performance. Participants completed two lexical decision tasks (LDTs) in which they decided as quickly as possible whether a string of letters was a word or a non-word. Each task used a different word list that contained 60 stimuli: 15 non-words (e.g., akmow, virpest), 15 negative but non-hostile words (e.g., cancer, gross), 15 hostility-related words (e.g. angry, kill), and 15 neutral words (e.g., lamp, pavement; see Appendix B). The order of word presentation within each list was randomized within subjects and the order of list presentation was counterbalanced between subjects. Each trial began with an asterisk centered on the computer screen for 500 ms. The stimulus word appeared in the same location and remained until the participant indicated, using the keyboard, whether the stimuli is a word or a non-word. A blank screen followed for 500 ms before the next trial began. Comparing mean response latencies of the aggressive words among conditions can reveal whether recognition of aggressive words is differentially facilitated, which is an implicit measure of anger. (Ayduk, Mischel, & Downey, 2002; Gollwitzer & Denzler, 2009).

Self-reported judgments and emotional reactions. Participants rated the fairness and moral wrongness of other players' actions during the game on 10-point Likert-type scales (Appendix C). They also rated their emotional reactions to the other players on 6-point Likert-type scales (Appendix C; see Appendix D for an overall timeline of the experimental session). Of major focus here were anger and envy toward unfair Decision-Makers and empathy toward victims. Our dependent measure of anger was the mean of three items ("angry," "mad," "outraged"; *alpha* = .90); envy was the mean of two items ("envious," "jealous"; *alpha* = .85); empathy was the mean of four

items (“compassionate,” “empathic,” “pity,” “sympathetic”; *alpha* = .86). See Appendix E for an ancillary measure (dominance) and analysis.

Chapter 3: Results

Descriptive Statistics

Means and standard deviations for all major variables appear in Table 1. Intercorrelations among all major variables appear in Table 2.

Analyses

Excluded participants. Twenty five participants revealed during the debriefing process that they had suspicious that either (a) the note they received was fabricated or (b) they had not actually interacted with real people. These participants were excluded from all analyses (total recruited $N = 237$; analyses $N = 212$). The number of participants excluded did not vary by condition , χ^2 (3, $N = 237$) = 2.35, $p = .502$, indicating that suspicions did not vary by condition.

Empathy manipulation check. Self-reported empathy toward victims varied significantly among the three witness conditions, $F(2, 159) = 3.52$, $p = .03$. Follow-up independent samples t-tests revealed that empathy in the no-instructions condition ($M = 3.05$, $SD = 1.22$) was higher than in the low-empathy condition ($M = 2.39$, $SD = 1.29$), $t(103) = 2.71$, $p = .01$, $d = .53$, and that empathy in the high-empathy condition ($M = 2.89$, $SD = 1.45$) was marginally significantly different from empathy in the low-empathy condition ($p = .07$), but not significantly different from the no-instructions ($p = .51$) condition (ds = .36 and .12, respectively; see Figure 1). Because the no-instructions and high-empathy conditions appeared to elicit approximately equal levels of empathy, these two conditions were combined for all subsequent analyses and will heretofore be referred to collectively as the "combined empathy" condition—which reported more empathy ($M = 2.97$, $SD = 1.34$) than the low-empathy condition, $t(160) = 2.57$, $p = .01$, $d = .44$.

Participants in the offset control condition did not witness unfairness. Instead, they were treated unfairly by the Decision-Maker. Therefore, their ratings of empathy toward the note-sender can function as a reference point from which to examine the nature of the empathy manipulation—that is, their ratings capture empathy felt as a result of reading the note without instructions and not witnessing the author of the note subsequently receiving harm. For participants in this "recipient of unfairness" condition, empathy for the author of the note (M = 2.5, SD = 1.64) was [marginally] significantly lower than it was for participants in the combined empathy condition (p = .06, d = -.31), and not significantly different from participants in the low-empathy condition (p = .71, d = .04). Thus, it appears that the effect of the empathy manipulation resulted from the combination of (a) reading the note either with no specific perspective-taking instructions or with explicit perspective-taking instructions, and then (b) observing the note-sender receive unfair treatment from the Decision-Maker. Without both of these conditions in place (i.e., among participants who either were instructed not to take the perspective of the note-sender, or participants who read the note without perspective-taking instructions but then did not go on to see the note-sender receive unfair treatment), empathy was low.

Punishment. A Kruskal-Wallis one-way ANOVA (distributions were non-normal) indicated that amount of punishment did not significantly vary by condition, H = .515, p = .77. To test whether witnesses of unfairness engaged in third-party punishment, one-sample Wilcoxon tests (with a hypothesized sample median of zero) were conducted on the punishment/reward distributions[4], revealing a significant amount of third-party

[4] Punishment, rewarding, and inaction were combined into a single punishment variable, such that amount (in $) punished took on positive values, amount rewarded took on negative values, and inaction took on a

punishment in both the combined empathy ($Z = 2.96$, $p < .01$, $N = 112$) and low-empathy ($Z = 2.51$, $p = .01$, $N = 50$) conditions (see Figure 2). Contrary to previous findings (Pedersen, McCullough, & Kurzban, 2012), there was not a significant amount of second-party punishment in the offset control condition ($Z = 1.39$, $p = .16$). Two possibilities may explain this lack of punishment: (a) it might have been driven by subjects who misunderstood that they had been harmed not by the author of the note, but rather, by the other of their two interaction partners[5]; and (b) there is some evidence that empathy or compassion toward one individual may reduce or eliminate punishment of another individual though the mechanisms underlying this possible phenomenon are unknown (Condon & DeSteno, 2011).

Self-reported and implicit anger. Self-reported anger toward Decision-Makers varied significantly among conditions, $F(2, 209) = 9.26$, $p < .01$. A planned linear contrast revealed that recipients of unfairness ($M = 1.69$, $SD = 1.22$) were angrier following Round 1 than were witnesses (combined empathy and low-empathy conditions combined; $M = 0.96$, $SD = 1.15$), $t(209) = 4.19$, $p < .01$, $d = .62$. A second planned linear contrast revealed that participants in the combined empathy condition ($M = 1.07$, $SD = 1.18$) trended toward reporting more anger than did those in the low-empathy condition ($M = .71$, $SD = 1.03$), $t(209) = 1.80$, $p = .07$, $d = .33$). Envy toward the Decision-Maker's ill-gotten gains accounted for witnesses' anger at unfairness in our previous study (see

---

value of zero. Thus, the one-sample Wilcoxon tests whether the observed sample median is significantly greater than a hypothesized median of zero (i.e., the distribution is significantly positive).

[5] Six out of 50 subjects in this condition *rewarded* the unfair Decision-Maker for treating them unfairly, three of whom gave their entire $5.00 endowment. Though nothing was revealed during debriefing that suggested confusion about the targets of their actions, some of these participants might have believed they were actually sending money to the person whose note they had read.

below; Pedersen et al., 2012). To control for the effects of envy, which was correlated with self-reported anger at $r(N = 212) = .37$, $p < .01$, I regressed anger on envy and saved the residuals. These residualized anger scores were used to measure self-reported anger in all subsequent analyses (except where otherwise noted). Controlling for envy, participants in the combined empathy condition ($M_{resid} = .21$, $SD_{resid} = 1.19$) were significantly angrier than those in the low-empathy condition ($M_{resid} = -.21$, $SD_{resid} = .91$), $t(121) = 2.48$, $p = .02$, $d = .40$ (Levene's test for equality of variances was significant, $F = 4.49$, $p = .04$, so degrees of freedom were adjusted from 160 to 121; see Figure 1)[6]. The measure of implicit anger (reaction times to hostility-related words in a lexical decision task) revealed no significant differences in anger between conditions, $F(2, 187) = 1.06$, $p = .35$[7].

Ratings of moral wrongness and fairness. Consistent with previous findings (Pedersen et al., 2012; Batson et al., 2007), participants in all four conditions did not differ in their ratings of how morally wrong they viewed the transgressor's behavior, $F(2, 209) = 1.25$, p = .29. However, contrary to previous findings, ratings of the fairness of the transgressor's behavior varied among conditions, $F(2, 209) = 4.05$, $p = .02$: Recipients of unfairness ($M = 3.00$, $SD = 2.37$) perceived the Decision Maker's behavior as more unfair

---

[6] For completeness: Participants in both the high-empathy ($M_{resid} = .15$, $SD_{resid} = 1.08$; $p = .06$) and no-instructions ($M_{resid} = .27$, $SD_{resid} = 1.30$; $p = .03$) conditions were angrier than those in the low-empathy condition, ds = .36 and .43, respectively.

[7] Data from 22 participants were excluded from analysis for one or more of the following reasons: all responses to nonwords were incorrect (indicating the participant was not properly doing the task), number of errors committed was ≥ 4 SD greater than the mean, number of outliers (classified as response times ≥ 3000ms) was ≥ 4 SD greater than the mean, computer error in writing the reaction time data. Remaining data were natural log-transformed to account for the common positive skew found in reaction time data (Ratcliff, 1993).

than did witnesses of unfairness (M = 4.04, SD = 2.23), planned linear contrast $t(209)$ = -2.83, p = .01, d = -.45. Participants in the combined empathy (M = 4.02, SD = 2.34) and low-empathy (M = 4.10, SD = 1.99) conditions did not differ in their ratings of fairness (p = .83, d = .04).

Comparisons with data from previous research. Importantly, and contrary to prediction, there was a greater-than-zero amount of punishment in the low-empathy condition. Thus, the low-empathy condition does not function as an ideal control condition with which to compare the effects of empathy and anger on punishment. Exploratory comparisons with data from two conditions (see Table 3 for summary statistics) from a previous study (Pedersen et al., 2012), though also not ideal (because in these group comparisons, participants were not randomly assigned to conditions, so the results are only quasi-experimental), provide some insight into the present findings.

No-empathy condition. One condition from our previous work, which I will call "no-empathy," was identical to the no-instructions condition in the present study, except that no notes were exchanged—that is, there was no elicitation of empathy (toward the Receiver): Participants merely witnessed the Receiver receive unfair treatment by the Decision-Maker. Participants in this condition did not engage in an amount of punishment significantly different from zero, Z = 1.48, p = .14, N = 65. Participants in the combined empathy condition (present study) did not punish significantly more than did participants in the no-empathy condition, Z = .01, p = .92, N = 177; they were, however, more empathic toward the Receiver, $t(175)$ = 2.52, p = .01, d = .22, and angrier toward the Decision-Maker, $t(175)$ = 2.06, p = .04, d = .33. Thus, the combined empathy condition in the present study did appear to increase empathy for victims of unfairness,

and anger toward unfair Decision-Makers, but it did not increase punishment of unfair Decision-Makers.

Witness of fairness condition. In the other condition, which I will call the "witness of fairness" condition, participants witnessed the Decision-Maker take nothing from the Receiver, and no notes were exchanged. In this condition, the median of the distribution of punishing/rewarding was significantly *less* than zero, $Z = -3.47$, $p < .01$, $N = 80$, indicating that participants typically rewarded fair Decision-Makers (a small amount, $M = \$0.34$). Participants in the combined empathy condition (present study) punished significantly more than did participants in the witness of fairness condition, $Z = 13.21$, $p < .01$, $N = 192$, and they were both angrier at the Decision-Maker, $t(144) = 5.11$, $p < .001$, $d = .69$ (Levene's test was significant, $F = 63.05$, $p < .01$, degrees of freedom adjusted from 190 to 144), and more empathic toward the Receiver, $t(190) = 10.69$, $p < .01$, $d = 1.59$.

The roles of empathy and envy in self-reported anger. In the previous study, significant anger at unfairness (relative to fairness) was completely accounted for by envy of the Decision-Maker's ill-gotten gains (that totaled \$9 to the participant's \$5; Pedersen, et al. 2012). The difference in anger between the combined empathy condition here and the "witness of fairness" condition from our previous study, however, cannot be completely explained in terms of envy: A one-way ANCOVA predicting (unresidualized) anger, with condition as a factor and envy entered as a covariate, showed a significant effect for condition, $F(1, 189) = 27.16$, $p < .01$, even when simultaneously controlling for the association of anger with self-reported envy, which itself was significant, $F(1, 189) = 10.27$, $p < .01$. When empathy was added to the model as a covariate, condition no longer

predicted anger, $F(1, 188) = .722$, $p = .40$, whereas both empathy, $F(1, 188) = 48.55$, $p < .01$, and envy, $F(1, 188) = 6.34$, $p = .01$, remained significant predictors[8]. Thus, it seems reasonable to conclude that, relative to participants who witnessed fairness in the previous study, participants in the present study who read a note about an individual who went on to be treated unfairly, and who were not restrained from feeling empathy for that (soon-to-be) victim by no-empathy instructions[9], experienced heightened anger toward the unfair Decision-Maker—apparently in part because of their heightened empathy for the victim—and were more inclined to impose a costly punishment.

---

[8] Additionally, a one-way ANCOVA predicting empathy, with condition as a factor and anger and envy entered as covariates, revealed significant main effects for condition, $F(1, 188) = 55.28$, $p < .01$, and anger, $F(1, 188) = 48.55$, $p < .01$. Given that empathy can account for differences in anger between groups, but anger cannot account for the differences in empathy between groups, it is reasonable to conclude that empathy likely plays a causal role in eliciting anger, rather than the reverse.

[9] Significant differences in anger between the low-empathy condition in the present study and the witness of fairness condition in the previous study, $F(1, 128) = 17.14$, $p < .01$, can be completely accounted for by envy toward the unfair Decision-Maker: In a one-way ANCOVA predicting anger, when condition was entered as a factor and envy entered as a covariate, the main effect for condition became [marginally] insignificant, $F(1, 127) = 3.76$, $p = .06$, and the main effect for envy was significant, $F(1, 127) = 38.44$, $p < .01$. When empathy was added to this model as a covariate (in addition to envy), it also significantly predicted anger, $F(1, 126) = 6.52$, $p = .01$, but its effect size was significantly lower than that of envy (partial $\eta^2 = .05$ and .17, respectively). Thus, whereas third-party punishment in the combined empathy condition appears be attributable mostly to empathic anger, third-party punishment in the low-empathy condition appears to be mostly attributable to *envious* anger.

Chapter 4: Discussion

Much of the extant literature on third-party punishment has been based on the group norm maintenance theoretical approach to human cooperation, which posits that humans punish social norm violators altruistically as third parties (e.g., Boyd et al., 2003; Fehr & Fischbacher, 2004; Fehr & Gächter, 2002; Gintis et al., 2003; Gintis, 2000; Henrich et al., 2006). Critically, both the evolutionary theorizing and the empirical findings that are marshaled in support of the claim for the existence of altruistic third-party punishment have been called into question (Burnham & Johnson, 2005; Hagen & Hammerstein, 2006; McCullough, Kurzban, & Tabak, in press; West et al., 2011). In a previous experiment that remedied experimental design problems with the third-party punishment game (Pedersen et al., 2012), we demonstrated that participants did not, in fact, punish altruistically on behalf of strangers; they did, however, readily punish those who had harmed them directly. Given the causal role that is often attributed to anger in the punishment literature (Nelissen & Zeelenberg, 2009; Pillutla & Murnighan, 1996; Sell, 2011; Srivastava & Espinoza, 2009) and our previous finding that recipients of unfairness became angry at transgressors and mere witnesses of unfairness did not, I hypothesized that the lack of third-party punishment on behalf of strangers in our previous work could be explained by witnesses' lack of anger toward transgressors. Furthermore, I hypothesized that their lack of anger toward transgressors was due to their lack of empathy for victims. The purpose of the present study was to investigate the proximate emotional underpinnings of third-party punishment—specifically, to test whether experimentally increasing empathy toward victims elicits anger and third-party

punishment. This hypothesis was generally supported, though comparisons with the previous study were helpful for interpreting the present results.

In the present study, subjects who read a note designed to elicit empathy on behalf of a victim reported more empathy than did those who read the note after receiving instructions designed to suppress empathy, and marginally more empathy than those in an offset control group consisting of people who read the note about a third party, but themselves were harmed by an unfair Decision-Maker. Self-reported anger significantly varied among conditions, such that recipients of unfairness were angrier than were participants in all three "witnesses of unfairness" conditions, and subjects in the combined empathy condition were angrier than those in the low-empathy condition. Results from the lexical decision task (our implicit measure of anger) did not indicate differences in anger. However, the significant effect found with the lexical decision task in our previous study had a small effect size (Pedersen et al., 2012), suggesting that the measure may be much less sensitive to changes in anger than our self-report measure. Taken together, these results generally support the hypothesis that increases in empathy for victims lead to increases in anger at transgressors. However, the increases in empathy and anger in the current study were not sufficient to increase third-party punishment: There was a significant amount of third-party punishment in the combined empathy and low-empathy conditions but the amount did not differ between the two groups. Moreover, for reasons that are difficult to explain, participants who were the recipients of unfairness did not significantly punish the individual who treated them unfairly. It is worthwhile to note that another research group has also found that experiencing empathy for one

individual reduced second-party punishment for another individual (Condon & DeSteno, 2011), so this result might actually be robust.

Comparing the data obtained in the present experiment with data from a previous experiment helped to shed additional light on the roles of empathy and anger in third-party punishment, although the caveats associated with causal interpretations of quasi-experimental data apply here. In the previous study (Pedersen et al., 2012), witnesses of unfairness (a no-empathy condition), as compared to witnesses of fairness, did not engage in a significant amount of third-party punishment and (when controlling for envy) did not report a significant amount of anger. Subjects in the combined empathy condition of the present study, however, did engage in a significant amount of third-party punishment and reported more anger (above and beyond what could be accounted for by envy) than did witnesses of fairness in the previous study. Importantly, the increase in anger not explained by envy could be completely explained by the increase in empathy, whereas the converse (anger explaining empathy) was not true. This pattern of results provides evidence that empathy might play a causal role in producing anger and third-party punishment, although future work that side-steps the inferential limitations associated with quasi-experimentation would help to strengthen this conclusion.

Interestingly, subjects in the combined empathy condition in the present study did not punish significantly more than did no-empathy witnesses of unfairness in the previous study[10]; they did, however, report more anger and empathy. Likewise, subjects in the no-

---

[10] It is important to note that subjects in the empathy condition *did* impose a significant amount of third-party punishment on transgressors, whereas those in the no-empathy condition of the previous study *did not*. However, there was not a significant difference in punishment in a direct comparison of the two

empathy witnesses-of-unfairness condition of the previous study did not punish more than did witnesses of fairness, nor did they report more anger. Thus, the comparisons between these three conditions (combined empathy in the present study, no-empathy witnesses of unfairness in the previous study, and witnesses of fairness in the previous study) provided an important insight: the combination of *both* witnessing unfairness *and* experiencing a significant amount of empathy for the victim is evidently necessary to elicit anger and third-party punishment. Furthermore, the differences between these groups suggests that, even though empathy was artificially induced and resulted in some punishment, third-party punishment on behalf of strangers is certainly not a robust and widespread phenomenon; it seems not to be the default behavioral response to learning that one anonymous stranger has harmed another anonymous stranger. Interestingly, this default of inaction can seemingly be overcome when a rather small amount of individuating information (in the form of a note in which the soon-to-be-victim describes a recent personal misfortune) causes the witness to experience empathy for the victim.

Limitations

One of the limitations of the current study is that the empathy manipulation I used did not lead to an ideal control condition in which witnesses of unfairness were not empathic toward victims: Even participants in the low-empathy condition—presumably because they read the note about the soon-to-be-victim's welfare—reported being empathic to some degree. Although some tentative conclusions can be drawn from comparisons to previous results, without the ideal control condition present in the current

---

conditions, which is likely a result of the rather large amount of variance in the distributions due to their zero-inflation.

experiment, some caution is needed in concluding decisively that increases in empathy increased third-party punishment. However, this limitation does provide critical insight in two areas. First, in conjunction with the previous results, the current data suggest that third-party punishment is still rather rare and mild, even when empathy is experimentally increased. Second, because punishment did not vary between the combined empathy and low-empathy conditions (which did have differences in reported empathy), it seems that small increases in empathy are insufficient, in and of themselves, to cause third-party punishment—that is, larger increases in empathy, more in line with the difference between witnesses of fairness and the combined empathy condition, are necessary to produce substantial differences in anger and punishment.

Another limitation of the current study is that third-party punishment, while statistically significantly present, was still rather rare (20 of 112 [17.9%] subjects punished some amount in the empathy condition). Consequently, there may not have been enough variability in punishment in the sample to examine fully the relationships between punishment, anger, and empathy. One possible way to remedy this paucity of punishment in future work is to decrease the cost of punishment; in the current study, there was a 1:4 cost-to-punish ratio. Whereas this ratio is lower than the typical 1:3 used in third-party punishment research (McCullough et al., in press), decreasing it even further is expected to encourage more punishment. Even allowing punishment to be imposed for free in the laboratory (e.g., subjects could burn up to $5 of the Decision-Maker's funds at no cost to themselves), could potentially help shed light on the underlying psychological processes driving punishment.

Future directions

When compared to results from our previous experiment, the present study showed promising results. However, because it lacked an ideal control condition, the current experiment should be replicated with a proper control. Given that (a) the results suggest a rather large amount of empathy is required to elicit a substantial increase third-party punishment; and (b) the theoretical basis for positing empathy's causal role in punishment lies in welfare interdependence, a future experiment might fruitfully consider welfare interdependence as a central focus. This goal might be accomplished in two straightforward ways: (a) having participants witness someone with whom they share varying levels of welfare interdependence (e.g., social category: sibling, friend, stranger) receive unfair treatment; or (b) experimentally manipulating welfare interdependence by structuring experimental games so that participants develop a mutually beneficial relationship with a partner that nets them different levels of benefits analogous to different levels of social category (e.g., a partner that consistently returns an additional 20% in a trust game, as a stranger might do, versus one who returns an additional 50%, as a friend might do).

There would be two main advantages to such an experimental approach. First, such an approach would more directly test the hypothesis that empathy is an output of psychological mechanisms designed to produce anger and motivate third-party punishment in response to cues that one's fitness interests have been threatened. Second, this method might produce greater differences in empathy for victims based on the real (or manipulated) welfare interdependence between the participant and the victim (e.g.,

participants should feel significantly more empathy for a sibling or a friend than for a stranger; Cialdini, Brown, Lewis, Luce, & Neuberg, 1997).

Additionally, the finding that recipients of unfairness in the current study viewed the unfair Decision-Maker's action as more unfair than did witnesses of unfairness warrants further investigation, as this effect was not found in our previous study. I have no solid theoretical explanations for this finding, but three possibilities come to mind. First, perhaps reading an empathy-inducing note had a general priming effect that made subjects more aware when they personally were cheated. Second, if the lack of punishment in this condition was actually due to the empathy felt toward another individual, perhaps the suppression of punishment resulted in subjects feeling more unfairness. Third, it is possible that this unpredicted finding is simply the result of Type I error. Regardless, this effect should certainly be further investigated in the future.

Conclusion

Herein I proposed a model of third-party punishment based on adaptationist principles that provides an alternative to the group norm maintenance model (Boyd et al., 2003; Fehr & Fischbacher, 2004; Fehr & Gächter, 2002; Gintis et al., 2003; Gintis, 2000; Henrich et al., 2006) of an altruistic benefit-delivery system: Psychological mechanisms that produce third-party punishment do so as a form of "revenge by proxy" whose function is to acquire inclusive fitness benefits for the punisher by deterring aggressors from harming individuals with whom the punisher has a vested fitness interest. Furthermore, based on principles of inclusive fitness and kin selection, I posited that empathy for victims as the output of a motivational system that computes the interdependence of the welfare of a victim and that of a prospective punisher. Should the

computed welfare interdependence be sufficiently high, I hypothesized, witnesses of unfairness will feel empathy for victims, and anger and punishment will consequently be directed toward the transgressor. That is, I proposed that empathy and anger are the proximate mechanisms that produce third-party punishment, and that they should only be experienced when one's welfare interdependence with a victim is sufficiently high. The present study provides a key first step in testing this model as the results suggest support for the proximate roles of empathy and anger in third-party punishment, and also poses some fascinating puzzles that can drive future experiments on this topic.

References

Aronson, E., Ellsworth, P. C., Carlsmith, J. M., & Gonzales, M. H. (1990). *Methods of research in social psychology*. New York: McGraw-Hill.

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*, 726-735.

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

Ayduk, O., Mischel, W., & Downey, G. (2002). Attentional mechanisms linking rejection to hostile reactivity: The role of "hot" versus "cool" focus. *Psychological Science, 13*, 443-448.

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

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

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science, 328*, 617-620.

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

Boyd, R, & Richerson, P J. (1988). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology, 132*, 337-56.

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

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

Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry, 37,* 47-59.

Carpenter, J. P., & Matthews, P. H. (in press). Norm Enforcement: Anger, indignation or reciprocity? *Journal of the European Economic Association*.

Carpenter, J. P., Verhoogen, E., & Burks, S. (2005). The effects of stakes in distribution experiments. *Economics Letters, 86*, 393-398.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997) Reinterpreting the empathy-altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology, 73,* 481-494.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature. 373*, 209-216.

Clutton-Brock, T. H. (2009). Cooperation between non-kin in animal societies. *Nature, 462*, 51-57.

Condon, P., & DeSteno, D. (2011). Compassion for one reduces punishment for another. *Journal of Experimental Social Psychology, 47*, 698-701.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187-276.

Cosmides, L, & Tooby, J. (1989). Evolutionary psychology and the generation of culture, part II. Case study: A computational theory of social exchange. *Ethology and Sociobiology, 10*, 51-97.

Engel, C. (in press). Dictator Games: A Meta Study. *Experimental Economics*.

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

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

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

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*, 980-994.

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

Fehr, E., & Henrich, J. (2003). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In P. Hammerstein (Ed.), *The genetic and cultural evolution of cooperation* (pp. 55-82). Cambridge, MA: MIT Press.

Gardner, A., & West, S. A. (2004). Cooperation and punishment, especially in humans. *The American Naturalist, 164*, 753-764.

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

Gintis, H, Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*, 153-172.

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology, 45*, 840-844.

Grafen, A. (1990). Do animals really recognize kin? *Animal Behaviour, 39*, 42-54.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*, 367-388.

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

Hamilton, W. D. (1963). The Evolution of Altruistic Behavior. *The American Naturalist, 97*, 354-356.

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

Henrich, J, & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior, 19*, 215-241.

Henrich, J, & Boyd, R. (2001). Why people punish defectors. *Journal of Theoretical Biology, 208*, 79-89.

Henrich, J, Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*, 795-815.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science, 312*, 1767-70.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Phil. Trans. R. Soc. B, 365*, 2635-2650.

Jensen, K., & Tomasello, M (2010) *Punishment*. Encyclopedia of Animal Behavior, Editors: Breed, MD, Moore, J, vol. 2, Academic Press (Oxford).

Kaplan, H., Hill, K., & Hurtado, A.M. (1990). Risk, foraging and food sharing among the Ache. In E. Cashdan (Ed.), *Risk and uncertainty in tribal and peasant economies.* Boulder: Westview Press.

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology, 5*, 289-305.

Lieberman, D., Tooby, J, & Cosmides, L. (2003). Does morality have a biological basis? An empirical test of the factors governing moral sentiments relating to incest. *Proc. R. Soc. B, 270*, 819-826.

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

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2011). Evolved mechanisms for revenge and forgiveness. In P. R. Shaver & M. Mikulincer (Eds.), *Human aggression and violence: Causes, manifestations, and consequences* (pp. 221–239). Washington, DC: American Psychological Association.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (in press). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*.

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

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Oosterbeek, H., Sloof, R., & Kuilen, G. van de. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics, 7*, 171-188.

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

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2012) Humans do not punish altruistically but they think they would. *Unpublished Manuscript.*

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

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

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510-532.

Richerson, P.J., & Boyd, R. (1997). The evolution of human ultra-sociality. In I. Eibl-Eibesfeldt & F. K. Salter (Eds.), *Ideology, Warfare, and Indoctrinability*. (pp. 71-96). New York: Berghahn Books.

Sell, A. N. (in press). The recalibrational theory and violent anger. *Aggression and Violent Behavior.*

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, 106*, 15073-15078.

Smith, J. E., Van Horn, R. C., Powning, K. S., Cole, A. R., Graham, K. E., Memenis, S. K., et al. (2010). Evolutionary forces favoring intragroup coalitions among spotted hyenas and other animals. *Behavioral Ecology, 21*, 284-303.

Srivastava, J., & Espinoza, F. (2009). Coupling and decoupling of unfairness and anger in ultimatum bargaining. *Journal of Behavioral Decision Making, 489*, 475-489.

Tooby, J., Cosmides, L., & Price, M. E. (2006). Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. *Managerial and Decision Economics, 27*, 103-129.

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

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77*, 273-295.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology, 20*, 415-432.

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

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

Table 1

Summary statistics of major study variables

	Overall		Recipient		Empathy		Low-Empathy	
Variable (scale)	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
$ Punished (-5 to 5)	0.43	1.65	0.50	2.41	0.37	1.33	0.52	1.36
LDT RT* (n/a)	6.54	0.23	6.52	0.20	6.53	0.22	6.58	0.26
Moral wrongness (1-9)	4.68	2.37	5.14	2.51	4.55	2.39	4.50	2.15
Fairness (1-9)	3.80	2.30	3.00	2.37	4.02	2.34	4.10	1.99
Anger (1-5)	1.13	1.21	1.69	1.22	1.07	1.18	0.71	1.03
Empathy (1-5)	2.72	1.42	2.50	1.64	2.97	1.34	2.39	1.29
Envy (1-5)	1.34	1.46	1.82	1.49	1.13	1.36	1.32	1.34

**Response time to hostility-related words, ln-transformed ms*

Table 2

Intercorrelations among major study variables.

Variable	1	2	3	4	5	6	7
$ Punished ‡							
LDT RT	.040						
Moral wrongness	.094	.037					
Fairness	-.101	-.031	-.364**				
Anger	.164*	-.001	.291**	-.351**			
Empathy	.021	-.057	.138*	-.059	.345**		
Envy	.096	.132	.029	-.037	.374**	.180**	

‡ *Spearman rank correlations used for $ Punished variable*

Table 3

Summary statistics of relevant conditions from Pedersen et al., 2012

	No-Empathy		Witness of fairness	
Variable (scale)	*Mean*	*SD*	*Mean*	*SD*
$ Punished* (-5 to 5)	0.24	1.37	-0.34	1.00
Moral wrongness (1-9)	5.18	2.30	1.61	1.79
Fairness (1-9)	4.26	2.43	8.48	1.24
Anger (1-5)	0.84	1.07	0.17	.046
Empathy (1-5)	2.47	1.15	1.03	1.09
Envy (1-5)	1.48	1.27	0.36	0.76

** Negative values indicate rewarding*
***Response time to hostility-related words, ln-transformed ms*

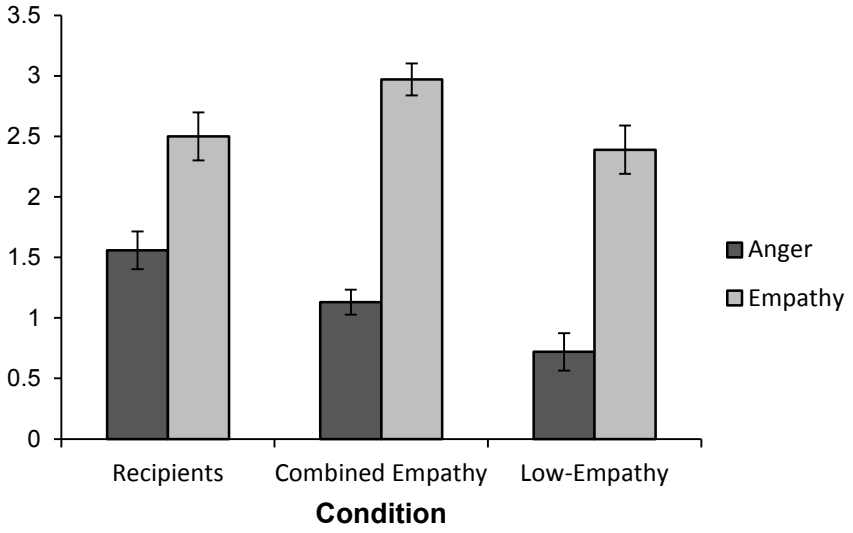


Figure 1. Self-reported anger at Decision-Maker and empathy toward Note-Sender (scale: 0 to 5; error bars = +/- 1 SE). Anger is displayed as the estimated marginal mean when controlling for envy, rather than residualized anger as used in the text, to preserve its scale. Residualized anger was used in the text to allow for the computation of effect sizes *d*; the significant differences between conditions are equivalent under either method.

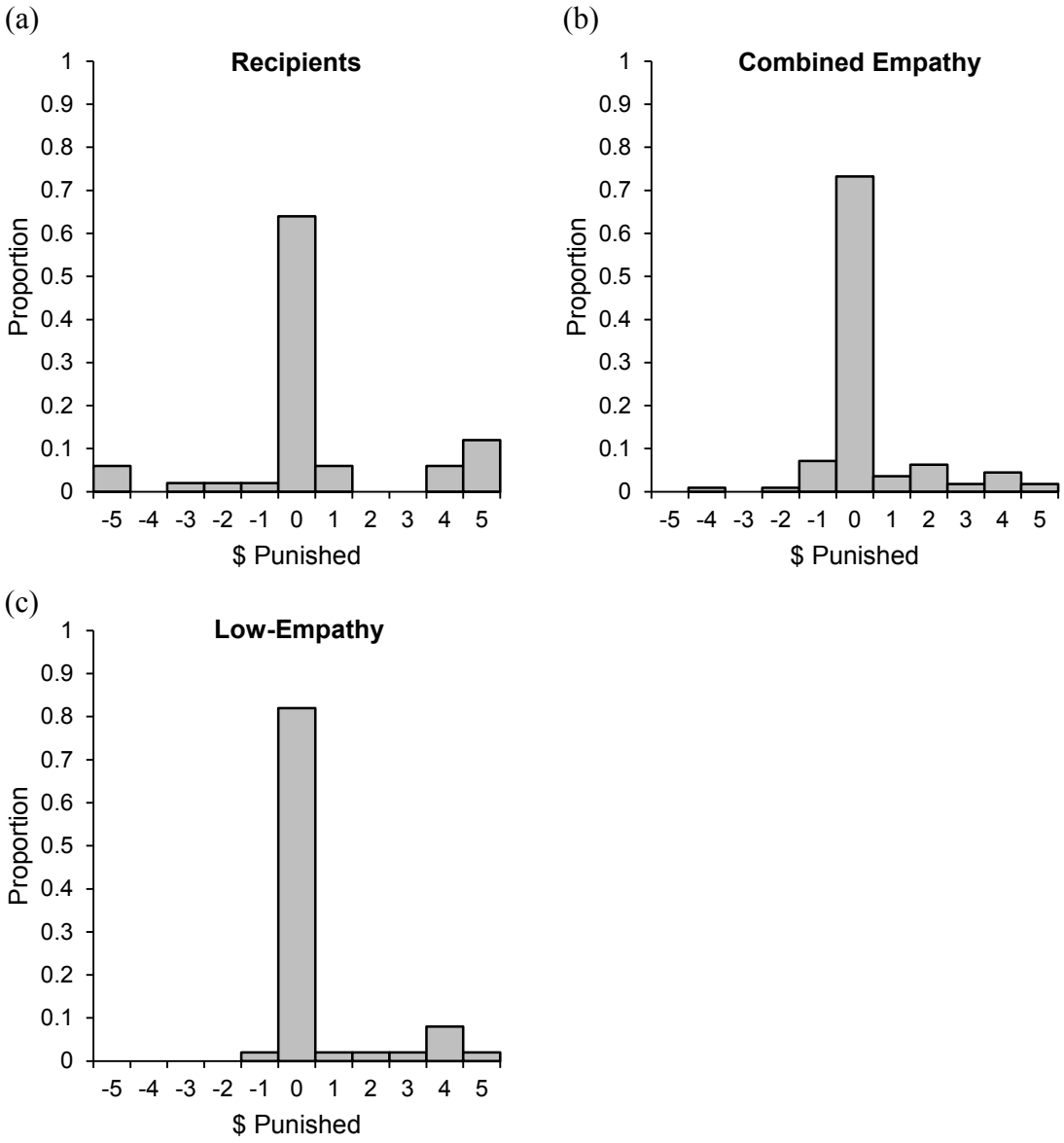


Figure 2. Punishment/reward distributions for (a) recipients; (b) combined empathy; and (c) low-empathy conditions. Negative values indicate amount (in $) participants rewarded the Decision-Maker; positive values indicate amount (in $) participants punished the Decision-Maker.

Appendix A

Empathy Manipulation

Low Empathy Instructions: While you are reading the note, try to take an objective perspective toward what is described. Try not to get caught up in how the Sender feels; just remain objective and detached.

High Empathy Instructions: While you are reading the note, try to imagine how the Sender feels about what is described. Try to imagine how it has affected the Sender's life and how he or she feels as a result.

Empathy Note: I'm supposed to write about something interesting thats happened to me lately. Well I don't know if this is interesting, but the only thing that I can seem to think of is that two days ago my boyfriend{girlfriend} broke up with me. We've been dating since our junior year in high school and have been really close and its been great being at UM together. I thought he{she} felt the same way but I guess that things have changed. Now he{she} wants to date other people. He{she} says that he{she} still cares a lot about me, but he{she} doesn't want to be tied down to just one person. I've been kind of upset. Its all I think about. My friends all tell me that I'll meet other guys{girls} and all I need is for something good to happen to cheer me up. I guess their right but so far that hasn't happened.

From (Batson, et al., 2007), edited to add grammatical errors.

Appendix B

Lexical Decision Task Words

Hostile	Negative	Nonword	Neutral
aggressive	accident	akmow	automobile
anger	cancer	arsgay	bookcase
attack	cockroach	avteal	cab
curse	crash	baxpov	cement
destroy	decayed	cawteg	chalk
enraged	dirty	cowmint	curtain
fierce	disaster	craffid	dress
fight	failure	dakewelt	form
fist	feces	dorsar	lamp
fury	fungus	fokyom	manual
hate	garbage	hasone	notebook
hit	gross	hunsop	pavement
hostile	infect	jalfig	room
kick	insect	jomtike	software
mad	itch	ligtuid	thread
mutilate	nausea	maigwen	
provoke	obese	makpak	
punch	pity	moufwent	
quarrel	pollute	naylim	
rage	poverty	nulhut	
revenge	puke	olpand	
scream	putrid	sekpair	
slap	rat	stawus	
spite	sickness	suzzle	
struggle	smog	virpest	
temper	stale	wimpow	
vengeance	trash	wodince	
vicious	virus	wongract	
violence	vomit	yaskog	
wrath	wart	zurpime	

(Ayduk, et al., 2002)

Appendix C

Self-Report Measures

How fair was the Decision-Maker's behavior toward the Receiver?

1	2	3	4	5	6	7	8	9
Not at all Fair								Totally Fair

How morally wrong was the Decision-Maker's behavior toward the Receiver?

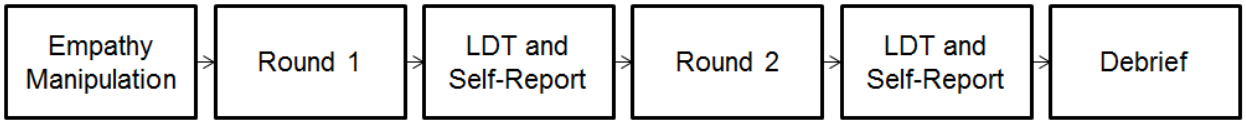
1	2	3	4	5	6	7	8	9
Not at all Morally Wrong								Totally Morally Wrong

Please indicate the extent to which you are feeling the following emotional response towards the Decision-Maker{Receiver}: _____

Angry	Mad
Annoyed	Offended
Bitter	Outraged
Compassionate	Pity
Content	Resentful
Empathic	Satisfied
Envious	Sympathetic
Grateful	Thankful
Happy	Vengeful
Irritated	Vindictive
Jealous	Warm

Appendix D

Timeline of Experimental Session



Appendix E

Ancillary Analysis

Dominance. The Assured-Dominant subscale (8-items; *alpha* = .76) from the Revised Set of Interpersonal Adjective Scales (IAS-R; Wiggins, Trapnell, & Phillips) was used to explore whether dominance was related to empathy and anger in response to unfairness. Subjects rated how much adjectives such as "dominant" and "forceful" applied to them on a Likert-type scale from 1 (*strongly disagree*) to 5 (*strongly agree*). Overall, there were no significant correlations with any of the major study variables. After splitting the data by condition, however, there was a significant correlation between dominance and ratings of moral wrongness in the combined empathy condition, $r = -.21$, $p = .02$.