2015-07-24

A Welfare Interdependence Approach to Third-Party Punishment

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

UNIVERSITY OF MIAMI

A WELFARE INTERDEPENDENCE APPROACH TO THIRD-PARTY PUNISHMENT

By

Eric J. Pedersen

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

August 2015

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

A WELFARE INTERDEPENDENCE APPROACH TO THIRD-PARTY PUNISHMENT

Eric J. Pedersen

Approved:

_____ Michael E. McCullough, Ph.D. Professor of Psychology	_____ Debra Lieberman, Ph.D. Associate Professor of Psychology
_____ J. Albert C. Uy, Ph.D. Associate Professor of Biology and Aresty Chair in Tropical Ecology	_____ William A. Searcy, Ph.D. Maytag Professor of Biology
_____ Blaine J. Fowers, Ph.D. Professor of Education and Psychological Studies	_____ Dean of the Graduate School

PEDERSEN, ERIC J. (Ph.D., Psychology)
A Welfare Interdependence Approach to Third-Party Punishment (August 2015)

Abstract of a doctoral essay at the University of Miami.

Doctoral essay supervised by Professor Michael McCullough.
No. of pages in text. (42)

Third-party punishment—the targeted infliction of costs on behalf of another by an unaffected third party—has been demonstrated in several species, including humans. Here I propose that one of the functions of third-party punishment is to deter future harm to victims with whom the punisher's welfare is interdependent. Additionally, I propose that this function is governed by psychological mechanisms that use internal regulatory variables called welfare trade-off ratios (WTRs) to guide social behavior via their outputs to motivational systems. Specifically, I propose that WTRs are used by the psychological mechanisms that regulate whether witnesses become angry in response to harms imposed on others, and thus, that they are key components of the system(s) that regulate third-party punishment. The goal of this dissertation was to test the causal role of welfare interdependence in third-party punishment by manipulating two WTR-relevant cues that were expected to raise subjects' WTR toward a partner who was initially a stranger, and then creating a situation in which the partner was harmed by another stranger, followed by an opportunity for the subject to punish the transgressor. In a laboratory experiment with 250 subjects, neither manipulation significantly affected subjects' WTRs for their partners. However, a noteworthy finding from this experiment is that there was not a significant amount of third-party punishment, which adds to a growing body of evidence suggesting third-party punishment on behalf of strangers is rare.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

Third-party punishment—the targeted infliction of costs on behalf of another by an unaffected third party—has been demonstrated in several species, including humans (Fehr & Fischbacher, 2004; Jensen, 2010; Konishi & Ohtsubo, 2015; Kurzban, DeScioli, & O'Brien, 2007; Lieberman & Linke, 2007; Raihani, Grutter, & Bshary, 2010; Smith et al., 2010). Punishment is inherently costly to the punisher because it requires time, energy, and potential exposure to counter-aggression. Hence, any theory that attempts to explain the existence of third-party punishment as a species-typical behavioral propensity—that is, the functional output of one or more evolved psychological mechanisms—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 (West, Griffin, & Gardner, 2007). On the basis of such reasoning, the use of third-party punishment should be expected to be selectively employed in situations in which, on average, the lifetime cost of punishment is outweighed by the lifetime benefits to the punisher. Here I propose that one of the functions of third-party punishment is to deter future harm to victims with whom the punisher's welfare is interdependent. Additionally, I propose that this function is governed by psychological mechanisms that use internal regulatory variables called welfare trade-off ratios to guide social behavior via their outputs to motivational systems (Sell, Tooby, & Cosmides, 2009; Tooby & Cosmides, 2008; Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008). Specifically, I propose that welfare trade-off ratios are used by the psychological mechanisms that regulate whether witnesses become angry in

response to harms imposed on others, and thus, that they are key components of the system(s) that regulate third-party punishment.

Is third-party punishment "altruistic?"

Much of the research on humans has conceptualized third-party punishment as a behavior that altruistically creates a benefit for the victim on whose behalf the third-party punisher intervenes (Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Fischbacher, 2003, 2004; Fehr & Gächter, 2002; Henrich et al., 2005). On this view, third-party punishment has been posited to result from a species-typical propensity to intervene and punish transgressors at a personal cost, *even when there is no possibility for direct or indirect benefits from the punishment*[1]. Due to the net costs "altruistic" punishers purportedly incur, some form of multi-level selection is often invoked to explain how such a behavior could have evolved (for review, see West, El Mouden, & Gardner, 2011). Though this proposal that third-party punishment is altruistic has been extremely influential across the social sciences, it has been challenged on both theoretical (Burnham & Johnson, 2005; Hagen & Hammerstein, 2006; McCullough, Kurzban, & Tabak, 2013; West et al., 2011) and empirical grounds (Krasnow, Cosmides, Pedersen, & Tooby, 2012; Pedersen, Kurzban, & McCullough, 2013).

The strongest empirical support for the altruistic punishment hypothesis comes from the third-party punishment game, which is a modified version of the dictator game (Fehr & Fischbacher, 2004). The game consists of a "Dictator" who chooses to give any

---

[1] Note that the term "altruistic" in the punishment literature has been used exclusively to refer to cases where indirect fitness benefits are not possible—hence, it does not refer to cases of punishment on behalf of kin that come at a net direct fitness cost to the punisher.

portion of a sum of money (or nothing at all) to a "Receiver," who has no influence on the interaction. A third player is the "Adjuster" and is instructed that they may pay a cost to reduce the Dictator's earnings following the Dictator's decision. Importantly, the Adjuster's own earnings are completely unaffected by the decision of the Dictator—the only way the Adjuster's monetary outcome can be affected is if he or she decides to punish the Dictator. Typically, adjusters in the third-party punishment game incur a personal cost to punish Dictators for unfair (i.e., less than 50% given to the recipient) splits of the money, despite deriving no financial benefit from doing so. For example, in the original third-party punishment game experiment approximately 60% of subjects punished unfair Dictator decisions, and the amount Adjustors punished Dictators was proportional to the unfairness of the split (Fehr & Fischbacher, 2004). These general results have been replicated many times in several cultures (Henrich et al., 2005; Henrich et al., 2010; Henrich et al., 2006; Marlowe et al., 2008)—though there is at least one exception (Marlowe, 2009)—and have been interpreted as evidence for the altruistic punishment hypothesis.

However, there are several methodological features of the standard third-party punishment game that, together, likely produce substantial experimental demand for punishment (Weber & Cook, 1972), which calls into question whether results from the game can be interpreted at face value (Pedersen et al., 2013). Two features stand out in particular. First, Adjusters in the game are presented only with the option to punish the Dictator or do nothing; they are not presented with the option for rewarding the dictator. Thus, (a) it is likely obvious to subjects that the experiment is about punishment; (b) if the subject wishes to take any action, his or her only option is to punish; and (c) if the

"true" average amount of punishment is actually zero, measurement error will be biased in the direction of punishment. Since the standard null hypothesis for the game is that no punishment will occur (e.g., Fehr & Fischbacher, 2004), even slight deviations from zero punishment can be deemed as support for the altruistic punishment hypothesis. Second, Adjusters' decisions in the typical game are made using the so-called strategy method (Selten, 1967), in which Adjusters are asked to respond to every possible Dictator choice *before the Dictators' decisions are actually revealed*. Hence, Adjusters must forecast how they would feel and respond for every possible scenario, rather than responding in real time to an actual decision. This is a critical issue in light of the fact that a substantial body of research shows that humans are notoriously inaccurate when forecasting their emotional and behavioral reactions to situations (Cook & Yamagishi, 2008; Kawakami, Dunn, Karmali, & Dovidio, 2009; Pedersen et al., 2013; Wilson & Gilbert, 2005).

To address these methodological issues, Pedersen et al. (2013) created a modified version of the third-party punishment game in which (a) Adjusters were given the opportunity to reward Dictators, as well as to punish or do nothing, and (b) Adjusters made a single decision *after* the Dictator's choice was revealed. In the modified game, Adjusters who witnessed Dictators treat Receivers unfairly did not punish Dictators, on average, amounts significantly different from zero. Additionally, after controlling for self-reported envy (which was necessary because unfair Dictators ended the game with more money that Adjusters), Adjusters who witnessed unfairness did not report that they were any more angry toward the Dictator than Adjusters who had witnessed fairness. In contrast, subjects who were personally treated unfairly by the Dictator did administer a significant amount of punishment and, relative to controls, reported a significant amount

of anger (when controlling for envy). Hence, in a modified version of the third-party punishment game, which was designed to reduce experimental demand for punishment, there was no evidence for altruistic third-party punishment. Additionally, In a separate experiment employing different methods, Krasnow et al. (2012) found that subjects' decisions to punish unfair splits in a trust game were not at all influenced by how the transgressor had treated other players in the past—the only predictor of punishment was how the transgressor had treated subjects themselves. Taken together, these recent results cast doubt on the hypothesis that the function of third-party punishment is to altruistically deliver benefits to strangers.

An alternative approach

Though it increasingly appears that humans do not punish "altruistically" as third parties, people *do* engage in third-party punishment (e.g., Phillips & Cooney, 2005). Largely overlooked in the human cooperation literature are accounts of third-party punishment based on the standard inclusive fitness maximization view of natural selection, which explains all known cases of third-party punishment in non-human animals (e.g., Raihani et al., 2010; Smith et al., 2010). Indeed, there are multiple plausible pathways for third-party punishment in humans to provide either direct or indirect fitness benefits to punishers, including reputational benefits (Barclay, 2006, 2013; Kurzban et al., 2007) or deterrence of future aggression directed toward the self (Delton, Krasnow, Cosmides, & Tooby, under review). In addition to these possibilities, I propose that one of the fundamental functions of third-party punishment (i.e., one of the primary social effects that caused third-party punishers' inclusive fitness to increase as the cognitive system or systems that regulate third-party punishment were evolving) is to

deter aggressors from harming individuals with whom the punisher has an inclusive fitness interest. On this view, the costs of third-party punishment can be offset via fitness benefits gained by deterring future harm toward victims whose welfare punishers perceive to be interdependent with their own (e.g., kin, mates, affines, friends, and coalition members). Because a third party will incur an indirect cost, via the cost imposed upon a victim by an aggressor, proportional to the extent that the third party's welfare is interdependent with the victim's, I anticipate a positive association of a potential third-party punisher's perceived welfare interdependence with the victim and his or her likelihood of engaging in third-party punishment on the victim's behalf. Likewise, because a third-party will also accrue an indirect benefit, via the benefit gained by the aggressor in imposing costs on the victim, proportional to the extent that the third party's welfare is interdependent with the aggressor's, I further anticipate a negative association of a potential third-party punisher's perceived welfare interdependence with the aggressor and his or her likelihood of engaging in third-party punishment on the victim's behalf.

Some existing evidence is consistent with these predictions. In a vignette study, Lieberman and Linke (2007) found that the "social category" (that is, the type of relationship between two people; e.g., family, friends) of both perpetrators and victims of crimes affected how much subjects reported that perpetrators should be punished, and how much time and energy subjects themselves would be willing to expend to bring a perpetrator to justice. Specifically, subjects reported that, for the same crime, (a) perpetrators who were family members should be punished less than either schoolmates or foreigners; (b) perpetrators who harmed a family member should be more harshly punished than those who harmed either schoolmates or foreigners; and (c) subjects

reported they would be willing to spend more days without pay and give up more weekends to help bring a perpetrator to justice who had harmed a family member than a perpetrator who had harmed either schoolmates or foreigners. (See also Bernhard, Fischbacher, and Fehr (2006) and Schiller, Baumgartner, and Knoch (2014) for qualitatively similar results of third-party punishment as a function of group membership).

In another study, Phillips and Cooney (2005) collected data on 136 recalled conflicts—which involved a total of 852 third-party witnesses—by interviewing 100 men imprisoned for assault or homicide. Of the third parties with "distant ties" (i.e., not a friend, family member, or fellow gang member) to one of the disputants, only 1% intervened in the conflict. In contrast, approximately 54% of third parties with "individual ties" (i.e., friends) and approximately 72% of third parties with "group ties" (i.e., family or gang members) to one of the disputants intervened.

Additionally, Pedersen & McCullough (2015) created a situation in which subjects either (a) witnessed a friend receive an insult from a stranger, (b) witnessed a stranger receive an insult from a stranger, or (c) received an insult from a stranger themselves. Later on in the session, under the guise of a different task, subjects were presented the opportunity to administer punitive sound blasts[2] to other subjects. Subjects

---

[2] There were two reasons for deviating from using an economic game for measuring punishment, and the same logic applies to why the same alternative method was chosen for the experiment reported in this dissertation. First, punishment distributions in economic games that do not use the "strategy method" can be very non-normal, requiring the use of lower-powered, non-parametric statistical tests (e.g., Pedersen et al., 2013). Use of the alternative sound blast measure resulted in normal distributions for which standard, parametric statistical tests were appropriate. Second, we wanted to use a method of punishment that was as cost-free as possible to encourage punishment (punishment tends to increase as its cost decreases;

readily punished insulters who had insulted either their friend or themselves, and results were mixed for punishing on behalf of a stranger. Additionally, subjects reported significantly more anger toward insulters who had mistreated their friend than they did toward insulters who had mistreated a stranger, and they reported significantly more empathy toward their friends that had received an insult than they did toward strangers who had received an insult.

In sum, existing evidence is consistent with the hypothesis that third-party punishment is preferentially implemented on behalf of people with whom punishers' welfare is interdependent. If this hypothesis is indeed correct, humans should possess psychological mechanisms that estimate their welfare interdependence with others. Without such computational mechanism(s), it is difficult to see how decisions to engage in third-party punishment could be made in an inclusive fitness-maximizing way[3]. For a simple example, consider the case of preferentially directing benefits toward kin. In order to take such action, an organism needs to, at a minimum, integrate two bits of information: (1) her relatedness to the potential target(s) of the benefits and (2) the costs/benefits associated with the action. Of course, the computations involved in such a

---

McCullough et al., 2013), so there would be enough variation in punishment to capture differences based on welfare interdependence.

[3] Of course, I do not mean to imply that the computational mechanisms involved in decisions to engage in third-party punishment measure or estimate inclusive fitness effects directly; rather, they will operate on cues that, ancestrally, were correlated with inclusive fitness effects. Additionally, I do not suggest that any of the computations I discuss occur consciously; rather, outputs of the computational mechanisms likely influence behavior via their effects on motivational systems. Lastly, "inclusive fitness-maximizing" does not refer to optimal decision-making; rather, it refers to the fact that, if the hypothesized computational mechanism(s) were indeed shaped by natural selection, they will appear as if they were designed to maximize their bearer's inclusive fitness. Hence, their outputs will lead to behavior that, *on average and under ancestral conditions*, led to increases in inclusive fitness.

system could be very rudimentary (e.g., both relatedness and cost/benefit could be simply considered as binary variables) and could be instantiated in many ways. Regardless of how the computation actually takes place, *some* integration of relatedness and costs/benefits must take place for benefits to be preferentially delivered to kin—that is, regardless of the actual instantiation, the systems that produce such behavior must be *functionally* computational. In the following sections, I outline one possible computational architecture for the psychological mechanisms underlying third-party punishment.

Sociality and welfare trade-offs

In social species, including humans, conflicts of interest between individuals are unavoidable—food eaten by one person is not available to another; a friend benefits from help that is costly to the helper; a mate courted by one person is, potentially, unavailable to another (Tooby et al., 2008). That is, actions taken by one individual often have fitness consequences for others and, to the extent that those fitness consequences for others impact the actor's own inclusive fitness, a selection pressure arises for taking into account others' welfare and estimating the extent to which the actor's own welfare is interdependent with others. Hence, sociality likely gives rise to computational systems that are capable of estimating welfare interdependence on the basis of multiple fitness-relevant inputs such that social behavior can be adaptively regulated (Roberts, 2005).

A proposed output of such a system is called a *welfare trade-off ratio* (WTR), which is hypothesized to be an internal regulatory variable that weights the welfare of another individual relative to the self and guides behavior accordingly through its effects on motivational systems (such as emotions; Tooby & Cosmides, 2008; Tooby et al.,

2008). WTR-generating mechanisms might compute WTR estimates for particular social partners on the basis of several inputs that are relevant to welfare interdependence. These might include ancestrally valid cues of genetic relatedness (e.g., sibship; Lieberman, Tooby, & Cosmides, 2007), past experiences of cooperative or exploitive interaction with the partners (Krasnow et al., 2012), estimates of the future interaction opportunities with the partners (Delton, Krasnow, Cosmides, & Tooby, 2011; McCullough, Pedersen, Tabak, & Carter, 2014), shared parental investments with the partners (Clutton-Brock, 1989), the partner's formidability and mate value (Sell et al., 2009), and mutual value (Tooby & Cosmides, 1996). Once a WTR estimate is established for a target individual, that estimate can be used to adaptively regulate social behavior[4].

On the basis of inclusive fitness thinking, we can expect that person *i* will take action when the "perceived" (see footnote 3) lifetime inclusive fitness benefits (b_i) of doing so outweigh the perceived lifetime inclusive fitness costs (c_i, which include opportunity costs associated with acting) of doing so, *on average*. That is, when:

$$(1) \qquad b_i > c_i$$

Additionally, the perceived benefit to *i* of an event that happens to *j* is its perceived benefit to *j*, discounted by *i*'s WTR for *j* (i.e., how much *i* values *j*'s welfare relative to *i*'s own):

$$(2) \qquad b_i = b_j * WTR_{i \rightarrow j}$$

---

[4] Note that WTRs will be updated whenever new, relevant inputs are processed. Estimates likely start off as diffuse priors, weighted by whatever WTR-relevant cues are available (e.g., mate value, formidability), and then are refined through Bayesian updating, with some cues (e.g., relatedness) having stronger impacts than others (e.g., a single interaction).

Likewise, the perceived cost to *i* of an event that happens to *j* is its perceived cost to *j,* discounted by *i*'s WTR for *j*:

$$(3) \qquad c_i = c_j * WTR_{i \rightarrow j}$$

Based on (1) through (3), we can formulate general rules for when social acts should be taken. According to Tooby et al. (2008), when person *i* encounters the opportunity to acquire a benefit (b_i) at a cost to person *j* (c_j), *i* should commit the act when the following inequality holds:

$$(4) \qquad b_i > c_j * WTR_{i \rightarrow j}$$

Re-arranging the terms in (4) based on (1) through (3), we can also see that when person *i* encounters the opportunity to provide a benefit to person *j* (b_j) at a personal cost to person *i* (c_i), *i* should commit the act when the following inequality holds:

$$(5) \qquad c_i < b_j * WTR_{i \rightarrow j}$$

When $WTR_{i \rightarrow j}$ is 0, *i* has no regard for *j*'s welfare (i.e., *i* would not incur any cost to benefit to *j*); when $WTR_{i \rightarrow j}$ is 1, *i* regards *j*'s welfare as equivalent to *i*'s own (i.e., *i* would incur any cost outweighed by the benefit to *j*). Note that when $WTR_{i \rightarrow j}$ equals the coefficient of relatedness, *r*, equation 5 reduces to Hamilton's rule (an inequality that states when altruistic behaviors can be favored by natural selection despite having a lifetime direct fitness cost to the actor; Hamilton, 1964; Tooby et al., 2008), which highlights the importance of WTR as a regulatory variable that integrates many factors: Although psychological mechanisms that produce behavior which satisfies Hamilton's rule will be favored by natural selection, mechanisms that can integrate relevant information in addition to estimates of relatedness will be more advantageous. For example, if a person has two siblings of equal relatedness but one sibling is an especially

generous cooperative partner whereas the other is exploitive, that person's welfare will be affected differently by the two siblings and behavior should be adjusted accordingly.

Now we consider a case in which person *i* (a third party) witnesses *j* directly incur a cost (c_j) by the action of a third person, *y*, who gains a benefit from the action (b_y). Person *i* has a WTR for person *j* ($WTR_{i \rightarrow j}$) and a WTR for person *y* ($WTR_{i \rightarrow y}$). Thus, even though *i* is not directly affected by the conflict, *i* incurs some of the cost to *j* and receives some of the benefit to *y* indirectly via *i*'s WTRs toward each party—that is, the cost and benefit are each weighted by the extent to which *i* values the welfare of the other person, relative to himself. Specifically, the indirect cost to *i* can be expressed as[5]:

$$(6) \quad c_i = c_j * WTR_{i \rightarrow j}$$

And the indirect benefit to *y* can be expressed as:

$$(7) \quad b_i = b_y * WTR_{i \rightarrow y}$$

Thus, the net cost[6] to *i* can be expressed as:

$$(8) \; c_{i(\text{net})} = (c_j * WTR_{i \rightarrow j}) - (b_y * WTR_{i \rightarrow y})$$

Given the structure of the human social environment (i.e., encountering a person once suggests a non-zero probability of reencounter; Krasnow, Delton, Tooby, & Cosmides, 2013), a net cost imposed on *i* implies that there is some non-zero probability of *y* imposing another cost in the future—hence, it may be beneficial for *i* to take action

---

[5] This may not be strictly true for aspects of WTRs that calibrated by factors other than relatedness. Rather, it may be the case that costs are only incurred indirectly *to the extent that they affect j's ability to provide benefits to i.* For example, if $WTR_{i \rightarrow j}$ is a function of how good of a cooperative partner *j* is for *i*, *j* could conceivably incur costs that do not impact the benefits he provides *i*. Despite this caveat, equations 6 and 7 should reflect general approximations for the indirect costs and benefits *i* receives.

[6] A negative value resulting from equation 8 would indicate a net benefit.

in an attempt to modify *y*'s perception of the costs and benefits associated with committing a similar act in the future.

Third party anger and punishment

Anger has been strongly implicated as one of the motivational systems involved in punishment in humans (Jensen, 2010; Petersen, Sell, Tooby, & Cosmides, 2010), and it has been proposed that anger's function is to motivate its bearer to employ bargaining tactics (including punishment) that resolve conflicts of interest to the benefit of the angry person (Petersen et al., 2010; Sell, 2011; Sell et al., 2009). Because anger can motivate any of a suite of bargaining tactics that are less costly than actual punishment (e.g., calling attention to the trangression, threats of punishment; Tooby et al., 2008), anger should only motivate punishment if a more cost-effective bargaining tactic is not available.

Based on equation 1, we should expect that *i* will only punish when the perceived benefits from doing so outweigh the perceived costs. There are likely many factors that contribute to estimates of costs and benefits associated with punishing, including: the likelihood the indirect harm will occur again; the likelihood that *y* will directly harm *i* in the future as implied by his treatment of *j*; the cost of punishment, including the likelihood of retaliation from *y*; the likelihood that punishment will recalibrate the behavior of *y*, and the associated benefits (or reduced costs); reputational effects; and direct deterrence effects. Because of the numerous factors that likely influence decisions to punish as a third party, a complete analysis of the specific conditions under which we would expect punishment would be extremely complex. However, because punishment is just one of a suite of tactics for bargaining for better treatment thought to be motivated by

anger, a focus on the conditions under which we would predict third parties to become angry at transgressors can shed light on general rules for minimum requirements necessary for punishment.

I propose that the anger system is triggered when a third party incurs a net cost as a result of a harm one person imposes on another—that is, person *i* will become angry at person *y* for harming person *j* when equation 8 yields a positive value. Hence, third parties will become angry when the cost to the victim, discounted by the third party's WTR toward the victim, exceeds the benefit to the harmdoer, discounted by the third party's WTR toward the harmdoer. Thus, for anger to be triggered and possibly lead to punishment, a third party needs both a sufficiently high WTR toward the victim and a sufficiently low WTR toward the harmdoer, relative to the costs and benefits incurred by each (i.e., to obtain a positive value for equation 8). If this hypothesis is correct, a third party's WTR toward a victim should be positively associated with anger toward the harmdoer, whereas the third party's WTR toward the harmdoer should be negatively associated with anger.

Preliminary evidence for the role of WTRs in producing third-party anger and punishment

Two studies to date have been conducted that provide preliminary, though mixed, support for the proposal that WTRs regulate third-party anger and punishment (additionally, see above in *An alternative approach* for consistent results from other studies that did not explicitly measure WTRs). First, Pedersen, McAuliffe, Shah, and McCullough (2015) asked subjects to "Please think of the last situation you can recall in which you witnessed someone attack, insult, or otherwise mistreat another person."

Subjects completed a WTR measurement (see Method below) toward each party, were asked to recall their emotional reactions toward each party, and were asked whether they intervened in the conflict in any way and, if so, what they did. Responses were coded as either "punishment" (cost inflicted on transgressor), "intervention" (intervened in any way), or "no involvement." WTR toward the victim was positively associated with anger toward the transgressor, as well as with both punishment and intervention. WTR toward the attacker was negatively associated with anger toward the transgressor, but was not associated with punishment or intervention.

Second, WTR measures were collected in the aforementioned experiment testing for third-party punishment on behalf of friends (Pedersen & McCullough, 2015). WTR toward the insulter was negatively associated with punishment but it was not associated with anger toward the insulter. Additionally, WTR toward the victim was not associated with anger toward the insulter or with punishment.

The present study: Do partner generosity and prospect of future interaction affect WTR, anger, and third-party punishment?

Although there multiple studies provide initial support for the hypothesis that third-party punishment is preferentially administered on behalf of those with whom the punisher's welfare is interdependent, and some mixed support for the hypothesis that punishment is regulated by WTRs, none of the studies mentioned have experimentally manipulated welfare interdependence, so we cannot draw firm conclusions regarding causality. The goal of this dissertation is to fill this gap and test the causal role of welfare interdependence on third-party punishment. To do so, I experimentally manipulated two WTR-relevant cues that are expected to raise subjects' WTR toward a partner (partner

generosity and probability of future interaction; see above in *Cooperation and welfare tradeoff ratios*) who is initially a stranger, and then I created a situation in which the partner was harmed by another stranger, followed by an opportunity for the subject to punish the transgressor.

Chapter 2: Method

Subjects

Subjects were 250 (136 female) undergraduate students at the University of Miami. I aimed to collect a sample size of at least 240 chosen based on estimated power exceeding .70 for main effects and interactions assuming medium effect sizes (i.e., $d \geq .5$).

Design

This experiment was a 3 (partner generosity: low, medium, high) by 2 (prospect of future cooperative interaction: low, high) between-subjects design.

Procedure

Subjects were run in small groups and led to believe that they were interacting with two other subjects over computers. In reality, they interacted with sham computer partners. See Figure 1 for a timeline of experiment procedures.

Subjects first played an iterated Trust Game (Berg, Dickhaut, & McCabe, 1995) with one of the sham partners. In the Trust Game, an "Investor" starts with an endowment of money and is given the chance to transfer some of it to the "Trustee." Transferred money is quadrupled, and the Trustee can then return some, none, or all of the proceeds back to the Investor. Subjects were "randomly assigned" to the Investor role, and their partner was assigned to the Trustee role, for three rounds of play (the number of rounds to be played was not specified in advance to avoid end of game effects). Subjects were given $1.50 to use in each round of the Trust Game, and any amount the subject transferred to the partner in each round was be multiplied by 4. Thus, a $1 transfer became $4 in the partner's account. After the Investor's transfer, the partner

could [ostensibly] back-transfer to the subject any amount (up to his total current holdings) he chooses.

Manipulation of perceived generosity. The partner's generosity was manipulated by having him return either 200% ("low"), 250% ("medium"), or 300% ("high") of the subject's investment in each of the three rounds (i.e., the partner always returned the same percentage).

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

Set-up for interpersonal harm. To create a situation in which an interpersonal harm could occur, this experiment used a multi-person extension of the "insulting essay evaluation" method (Harmon-Jones, Amodio, & Zinner, 2007; Harmon-Jones & Sigelman, 2001). Subjects were given five minutes to write (via PC) a short essay on a personally important issue. They were told that their (and, supposedly, the two sham subjects') essays would be submitted electronically and then circulated to the others, and each subject would read the other two subjects' essays and provide some very brief remarks to each of them. Subjects were also told that they would subsequently read the other subjects' evaluations of all of the essays.

WTR measurement. After submitting their own essays, and before reading the other ostensible participants' essays, subjects completed WTR measurements vis à vis the other subjects. The WTR measurement instructs subjects to make 10 binary decisions

about whether they would prefer to receive a certain amount of money for themselves (in descending order from $85 to $5 in $10 increments, and $0) or to confer a fixed amount ($75) to a focal individual (Jones & Rachlin, 2009)—in this case, each of the other two subjects. Using these data, each subject's point of indifference relative to the other two subjects was calculated (i.e., the midpoint between the last amount that the subject chooses to allocate to him/herself and the first amount at which the subject chooses to allocate $75 to the target) and divided by $75 to calculate a WTR. Although the rewards to be used here are hypothetical, people discount real and hypothetical rewards similarly (Madden, Begotka, Raiff, & Kastern, 2003). The order in which subjects completed the WTR evaluations for the two other people was randomized.

Interpersonal harm. Next, subjects read each of the other two subjects' evaluations of the essays. After reading each evaluation, subjects rated how "fair/unfair," and how "accurate/inaccurate" they thought it was on scales from 0 (not at all) to 9 (totally). A composite fairness/accuracy score was created by taking the mean of these two items (*alpha* = .87). This step was designed to ensure that subjects attended to the insulting evaluation and served as a manipulation check to confirm that the insult is perceived as unfair (relative to the other evaluations). The evaluations from the other subjects were bogus, and slightly positive (e.g., "I can understand why a person would think like this.") except for the stranger's evaluation of the subject's Trust Game partner's essay, which was negative: "I can't believe an educated person would think like this. I hope this person learns something while at UM". We have used this manipulation previously and have found that the insult elicits anger (Pedersen & McCullough, 2015), as have others (Harmon-Jones et al., 2007; Harmon-Jones & Sigelman, 2001).

Self-reported judgments and emotional reactions. Subjects then rated their emotional reactions to the other players on 6-point Likert-type scales from 0 (not at all) to 5 (extremely). Of major focus here was anger toward insulters (mixed among several distractors). A composite anger score was formed by taking the mean of participants' rating of how "angry," mad", and "outraged" they reported feeling toward the insulter (*alpha* = .92 in the present sample).

Dependent variable: Sound blast ("Punishment"). Next, subjects were told that the experimenters were evaluating various sounds for use in future experiments and needed some feedback on how pleasant or unpleasant the sounds were, and how this changed over the duration of a sound sample. Subjects were then told that they had been randomly chosen to be an "audio administrator" and to assign sound samples to the other subjects, who had been assigned to be sound raters. Subjects listened to three short samples, of different volumes, of an unpleasant static sound through headphones and were asked to rate how pleasant/unpleasant it was. Next, they assigned a volume level from 1 (quietest) to 10 (loudest), and held down the space bar (i.e., they assigned a duration by holding down the space bar) to play the noise for each of the other subjects individually. Subjects were be led to believe the sound was playing in real time for the other subject, but the subject did not hear the sound while holding down the space bar. As in Pedersen and McCullough (2015), a composite measure of punishment was created by taking the mean of the standardized values of both volume and duration (natural log-transformed due to skewness) of the sound blast. These two values were moderately correlated, $r(197) = .41$, $p < .001$, and yielded a composite whose internal consistency reliability was estimated at *alpha* = .58.

Predictions

I predicted main effects of both generosity and probability of future interaction on punishment of the insulter and self-reported anger toward the insulter, as well as main effects on WTR toward the subject's Trust Game partner. I also tested for interactions between both independent variables, but did not have a specific prediction. Additionally, I predicted a positive correlation between WTR toward the partner in the Trust Game and anger and punishment of the insulter; likewise, I predicted a negative correlation between WTR toward the insulter and anger and punishment of the insulter. If these predictions were supported, I planned test whether the main effects of generosity and probability of future interaction on anger and punishment are mediated by WTRs.

Chapter 3: Results

Descriptive Statistics

Means and standard deviations (overall and broken down by condition) for all major variables appear in Table 1. Correlations among all major variables appear in Table 2.

Analyses

Excluded participants. Forty-four participants revealed during the debriefing process that they had suspicions that either (a) some of the interactions in the experiment were fabricated or (b) they had not actually interacted with real people. These participants were excluded from all analyses (total recruited $N = 250$; analyses $N = 206$).

Manipulation check: Perceived fairness/accuracy of the insulting review. To test whether the insulting review was perceived as unfair (relative to the slightly positive review the victim sent the insulter), I conducted a full-factorial two-way repeated measures ANOVA predicting fairness/accuracy with reviewer (insulter; victim) as a within-subjects factor, and partner generosity (low, medium, high) and likelihood of future interaction (low, high) as between-subjects factors. The main effect for reviewer was significant, $F(1, 199) = 477.40$, $p < .001$, partial $\eta^2 = .71$, such that subjects reported the insulter's review ($M = 3.27$, $SD = 2.28$) as significantly less fair/accurate than the victim's review ($M = 7.50$, $SD = 1.61$), and none of the interaction terms were significant (*p*s from .302 to .765), indicating that this effect did not vary by condition.

Manipulation check: Welfare trade-off ratio (WTR) toward the partner. To test whether the manipulations of perceived welfare interdependence, as measured by WTR, were effective, I conducted a two-way ANOVA predicting WTR toward the

partner with partner generosity, prospect of future interaction, and their interaction as predictors. The model was not significant, $F(5, 161) = 1.19$, $p = .317$[7], and dropping the interaction term had no effect on the significance of the model, $F(3, 163) = 1.01$, $p = .389$. Hence, neither the manipulation of partner generosity, nor the manipulation of the possibility of future interaction, increased subjects' WTRs toward their partner (who would subsequently become the victim of an insult; see Figure 2.

To better understand the failure of the generosity manipulation, I ran one-way ANOVAs predicting (a) the total amount of money transferred to the partner in the three rounds of the trust game and (b) the total profit earned over the three rounds of the trust game. The total amount of money transferred did not vary by condition, $F(2, 203) = .161$, $p = .851$ (range: \$2.80 to \$2.92, of a possible \$4.50). However, the total profit *did* vary by condition, $F(2, 203) = 35.38$, $p < .001$: participants in the high-generosity condition ($M = \$5.71$, $SD = \$2.67$) earned significantly more money than those in the medium-generosity condition ($M = \$4.39$, $SD = \$1.93$; $t = 3.77$, $p < .001$, $d = .57$) who, in turn, earned more than those in the low-generosity condition ($M = \$2.80$, $SD = \$1.29$; $t = 4.51$, $p < .001$, $d = .97$). Thus, the manipulation was successful in creating partners that "valued" the subjects differentially given the same level of investment but these differences in value were apparently not strong enough—or there were not enough rounds in the game—to, in turn, lead to increases in subjects' valuations of the partners.

---

[7] The reduced degrees of freedom in this test resulted from the inability to calculate WTRs for some subjects because an indifference point could not be determined by their responses on the scale. All reductions in degrees of freedom in tests without WTRs resulted from missingness due to computer malfunction at some point during the experiment.

Self-reported anger toward the insulter. A two-way ANOVA predicting anger toward the insulter with partner generosity, prospect of future interaction, and their interaction as predictors was not significant, $F(5, 195) = 1.42$, $p = .218$ (see Figure 3). Dropping the interaction term led to a model that trended toward significance, $F(3, 197) = 2.34$, $p = .074$, in which likelihood of future interaction significantly predicted anger *opposite* the predicted direction, with subjects in the low likelihood condition ($M = 0.97$, $SD = 1.25$) reporting more anger than those in the high likelihood condition ($M = 0.66$, $SD = 0.95$) , $F(1, 199) = 4.09$, $p = .045$, partial $\eta^2 = .02$.

An OLS regression predicting anger toward the insulter with WTR toward the insulter and WTR toward the victim as predictors was not significant, $F(2, 154) = .43$, $p = .651$.

To evaluate the relative level of anger toward insulters versus victims of insults, I conducted a full-factorial two-way repeated measures ANOVA predicting anger with target (insulter; victim) as a within-subjects factor, and partner generosity and likelihood of future interaction a between-subjects factors. The main effect of target was significant, $F(1, 195) = 24.40$, $p < .00$, partial $\eta^2 = .11$ such that subjects reported significantly more anger toward insulters ($M = .82$, $SD = 1.12$) than toward victims ($M = .39$, $SD = .79$), and none of the interaction terms were significant (ps from .144 to .852). Hence, subjects reported significantly more anger toward insulters than toward victims and this effect did not vary by experimental condition.

Punishment. A two-way ANOVA predicting punishment of the insulter with partner generosity, prospect of future interaction, and their interaction as predictors was not significant, $F(5, 193) = 1.19$, $p = .313$, and dropping the interaction term had no

qualitative effect on the significance of the model, $F(3, 195) = 1.69$, $p = .171$. Hence, neither partner generosity nor likelihood of future interaction affected punishment of the insulter (see Figure 4).

An OLS regression predicting punishment of the insulter with WTR toward the insulter and WTR toward the victim as predictors trended toward significance, $F(2, 152) = 2.656$, $p = .073$. WTR toward the insulter significantly predicted punishment ($b = -.92$, $SE = .41$, $p = .027$, 95% CI = -1.74 to -.11, $\beta = -.37$) such that going from 0 (i.e., not at all valuing the insulter's welfare) to 1 (valuing the insulter's welfare as much as one's own) on the WTR scale led to an estimated .37 standard deviation reduction in punishment. WTR toward the victim did not significantly predict punishment, though the effect was in the predicted direction ($b = .71$, $SE = .42$, $p = .096$, 95% CI = -.12 to 1.54, $\beta = .28$). Thus, though the manipulations of victim WTR were not effective, there was still some evidence that WTR toward the attacker regulated punishment in the theoretically predicted direction[8].

To evaluate the relative level of punishment administered to insulters versus victims of insults, I conducted a full-factorial two-way repeated measures ANOVA predicting punishment with target (insulter, victim) as a within-subjects factor, and partner generosity and likelihood of future interaction as between-subjects factors. The main effect of target was not significant, $F(1, 193) = .00$, $p = .968$, nor were any of the

[8] Given that variance in WTRs was not attributable to the manipulation, it could be the case that this positive association resulted from a third variable that causes individual differences in baseline WTRs and likelihood of punishment—for example, strength and fighting ability might calibrate both, leading to the observed correlation.

interaction terms (*p*s from .255 to .986). That is, subjects did not punish insulters any more than they "punished" victims of the insults.

Given the lack of punishment of insulters relative to victims, I tested whether the significant negative association between WTR toward the insulter and punishment of the insulter may have arisen simply due to a general negative association between WTR and administering the sound blast measure (i.e., that it was unrelated to punishing the insult). In an OLS regression predicting punishment of the *victim* with WTR toward the insulter and WTR toward the victim as predictors, neither WTR toward the insulter (b = -.45, SE = .40, p = .26, 95% CI = -1.23 to .33, β = -.18) nor WTR toward the victim (b = .03, SE = .41, p = .94, 95% CI = -.76 to .83, β = .01) predicted punishment.

Chapter 4: Discussion

The explanatory and predictive power of the standard inclusive fitness maximization view of natural selection has been largely overlooked among researchers who use evolutionary theory as a heuristic for thinking about third-party punishment in humans. Using such an approach, here I proposed that the fundamental function of third-party punishment is to deter future harm to victims with whom the punisher's welfare is interdependent. Additionally, I proposed that this function is governed by psychological mechanisms that use internal regulatory variables called welfare trade-off ratios (WTRs) to guide social behavior. Specifically, I proposed that WTRs are used by the psychological mechanisms that regulate whether witnesses become angry in response to harms imposed on others, and thus, that they are key components of the system(s) that regulate third-party punishment.

Though multiple studies have provided initial support for the hypothesis that third-party punishment is preferentially administered on behalf of those with whom the punisher's welfare is interdependent (e.g., Lieberman & Linke, 2007; Pedersen & McCullough, 2015; Phillips & Cooney, 2005), and some mixed support for the hypothesis that punishment is regulated by WTRs (Pedersen et al., 2015), the causal role of welfare interdependence in third-party punishment has not been tested. The goal of this dissertation was to fill this gap by manipulating two WTR-relevant cues (partner generosity and probability of future interaction) that were expected to raise subjects' WTR toward a partner who was initially a stranger, and then creating a situation in which the partner was harmed by another stranger, followed by an opportunity for the subject to

punish the transgressor. Neither manipulation significantly affected subjects' WTRs for their partners.

Two possibilities seem likely to account for the failure of these manipulations. First, and most obvious, it may simply be the case that partner generosity and likelihood of future interaction do not, in fact, affect WTRs as predicted. Second, it could be the case that the manipulations were simply not strong enough. This latter explanation seems like a strong possibility for the generosity manipulation, in particular, given that subjects across conditions did not vary their level of transfers in the Trust Game in response to partner generosity—thus, it could have been the case that three rounds of the game were simply not enough, or the amounts of money at stake were not enough, to allow for enough feedback on the partner's value (see Future Directions, below).

Given the ineffectiveness of the generosity manipulation at creating between-condition differences in WTRs, it is unsurprising that there were also no significant effects of partner generosity on either punishment of the insulter or anger toward the insulter. However, WTR toward the insulter did predict punishment in the predicted negative direction and, though not statistically significant, the association between WTR toward the victim and punishment of the insulter was in the predicted direction. Neither WTR toward the victim nor WTR toward the insulter predicted anger toward the insulter, though the base rate of anger was very low (M = .82 on a scale from 0 to 5), possibly leading to range restriction issues. This pattern of results is similar to those in a previous third-party punishment experiment using a virtually identical experimental protocol without the attempted manipulation of WTRs (Pedersen & McCullough, 2015).

Though the likelihood of future interaction manipulation did not significantly predict punishment of the insulter, it did have a small effect on anger toward the insulter such that a high likelihood of future interaction with the victim led to *less* reported anger toward the insulter, contra my prediction. Whereas this result is opposite my prediction, the combination of the ineffectiveness of the manipulation in affecting WTR, the small effect size (~2% of variance explained), and the possibility of a false positive make it difficult to interpret on its own, so further work to investigate the role of the likelihood of future interaction in third-party anger and punishment is warranted

A particularly noteworthy finding from this experiment is that despite punishment being cost-free (as opposed to being costly in typical economic games), there was not a significantly greater amount of punishment directed toward insulters than toward the victims of the insults. Though "punishment" of victims is not a perfect control group, this finding adds to a growing body of evidence suggesting third-party punishment on behalf of strangers is rare (Krasnow et al., 2012; Pedersen et al., 2013; Pedersen et al., 2015; Pedersen & McCullough, 2015).

Limitations

The major limitation of this experiment was the failure to manipulate perceived welfare interdependence, as measured by WTR, via the partner generosity and likelihood of future interaction manipulations. Hence, the main aim of the experiment—to test the causal role of welfare interdependence on third-party punishment—was not accomplished. Additionally, 44 subjects reported suspicion about the legitimacy of some of the procedures or whether they were interacting with real people and thus had to be excluded from analyses, which reduced the number of useable data points to below my

goal of 240. Thus, the experiment was probably not ideally powered. However, if the population effect sizes are indeed close to zero, as these results would suggest (assuming the manipulations could not be improved), then no study would have adequate power.

Future Directions

Though the present experiment did not provide the necessary manipulation of welfare interdependence to properly test my hypotheses, the design could possibly be modified to achieve that aim. First, it seems as though there were not enough rounds in the Trust Game for subjects to refine their WTR estimates toward their partner enough to generate differences among conditions of the partner generosity manipulation. A greater number of rounds would likely reduce the variance in subject's WTRs and lead to distinctions among conditions, which could also be made more distinct by increasing the percentage differences in generosity between conditions. Furthermore, related to reducing variance in WTR estimates, a finer-grained WTR measure could be used—though this would present a tradeoff with increased length of the questionnaire, doubling the number of tradeoff decisions subjects make should substantially reduce the variability in their responses. Another possibility would be to create an adaptive scale in which subjects are presented with tradeoff decisions in between the points of the original scale that mark their switchpoint—for example, if a subject chooses to keep $55 for herself instead of $75 for the partner, but chooses $75 for the partner rather than $45 for herself, the scale could be refined between those two points to get a more accurate estimate of her WTR.

More generally, despite the shortcomings of the present research, a shift in focus from studying third-party punishment on behalf of complete strangers to punishment on behalf of those with whom the punisher's welfare is at least somewhat interdependent is

likely to yield a significant advancement in our understanding of the both the function and the psychological underpinnings of third-party punishment.

Conclusion

Herein I proposed and sought to test an account of third-party punishment in humans suggesting that function of third-party punishment is to deter future harm to victims with whom the punisher's welfare is interdependent. Though the failure of my experimental conditions in manipulating welfare interdependence make the largely null findings difficult to interpret, the present experiment does provide a useful starting point for designing future studies to more effectively test my hypotheses and is a key step in the direction of investigating explanations for third-party punishment grounded in an inclusive fitness theoretical framework.

References

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

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior, 34*(3), 164-175. doi: 10.1016/j.evolhumbehav.2013.02.002

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

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

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

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

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

Cook, K. S., & Yamagishi, T. (2008). A defense of deception on scientific grounds. *Social Psychology Quarterly*, 215-221.

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

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (under review). Inferences about risks of personal mistreatment correlate with third party punishment.

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

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

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

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

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

Harmon-Jones, E., Amodio, D. M., & Zinner, L. R. (2007). Social psychological methods of emotion elicitation. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 91-105). New York: Oxford.

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

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

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

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

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B, 365*, 2635-2650.

Jones, B. A., & Rachlin, H. (2009). Delay, probability, and social discounting in a public goods game. *Journal of the Experimental Analysis of Behavior, 91*, 61-73.

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

Konishi, N., & Ohtsubo, Y. (2015). Does dishonesty really invite third-party punishment? Results of a more stringent test. *Biology Letters, 11*(5), 20150172.

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

Krasnow, M. M., Delton, A. W., Tooby, J., & Cosmides, L. (2013). Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation. *Scientific Reports, 3*.

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

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. (2007). The architecture of human kin detection. *Nature, 445*, 727-731.

Madden, G. J., Begotka, A. M., Raiff, B. R., & Kastern, L. L. (2003). Delay discounting of real and hypothetical rewards. *Journal of Experimental and Clinical Psychopharmacology, 11*, 139-145.

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

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

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

McCullough, M. E., Pedersen, E. J., Tabak, B. A., & Carter, E. C. (2014). Conciliatory gestures promote forgiveness and reduce anger in humans. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1405072111

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

Pedersen, E. J., McAuliffe, W. B., Shah, Y., & McCullough, M. E. (2015). Do third parties punish outside of the lab? A field study. *Unpublished manuscript*.

Pedersen, E. J., & McCullough, M. E. (2015). Third parties engage in anger-motivated punishment on behalf of friends, but not strangers. *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. (2005). Aiding peace, abetting violence: Third parties and the management of conflict. *American Sociological Review, 70*, 334-354.

Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science, 327*, 171.

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

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior, 35*(3), 169-175.

Sell, A. (2011). The recalibrational theory and violent anger. *Aggressive Behavior, 16*, 381-389.

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

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

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

Tooby, J., & Cosmides, L. (1996). Friendship and the Banker's Paradox: Other pathways to the evolution of adaptations for altruism. In W. G. Runciman, J. M. Smith & R. I. M. Dunbar (Eds.), *Evolution of Social Behaviour Patterns in Primates and Man. Proceedings of the British Academy* (Vol. 88, pp. 119-143).

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

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

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

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

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.

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting knowing what to want. *Current Directions in Psychological Science, 14*(3), 131-134.

Table 1. *Descriptive statistics of major study variables*

Condition	Overall						
Future Interaction:		Low			High		
artner Generosity:		Low	Med	High	Low	Med	High
Variable	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Trust Game earnings ($)	4.31 (2.37)	2.85 (1.37)	4.15 (2.02)	5.46 (2.77)	2.73 (1.18)	4.58 (1.86)	5.92 (2.59)
Sent in Trust Game ($)	2.86 (1.30)	2.85 (1.37)	2.77 (1.35)	2.73 (1.39)	2.73 (1.78)	3.05 (1.24)	2.96 (1.29)
Fair/accurate (insulter)	3.27 (2.28)	3.30 (2.69)	3.02 (2.12)	3.61 (2.05)	2.78 (1.75)	3.26 (2.45)	3.47 (2.32)
Fair/accurate (victim)	7.5 (1.61)	7.86 (1.19)	7.50 (2.06)	7.92 (0.88)	7.23 (1.66)	7.41 (1.84)	7.01 (1.76)
WTR (insulter)	0.58 (0.36)	0.62 (0.35)	0.56 (0.35)	0.54 (0.39)	0.55 (0.38)	0.63 (0.35)	0.54 (0.37)
WTR (victim)	0.59 (0.35)	0.59 (0.32)	0.60 (0.34)	0.58 (0.37)	0.46 (0.38)	0.70 (0.32)	0.59 (0.35)
Anger (insulter)	0.82 (1.12)	1.02 (1.35)	1.09 (1.31)	0.82 (1.06)	0.63 (0.82)	0.85 (1.14)	0.48 (0.81)
Anger (victim)	0.39 (0.79)	0.46 (0.88)	0.40 (0.88)	0.40 (0.91)	0.46 (0.83)	0.38 (0.64)	0.22 (0.63)
Punishment (insulter)	0.00 (0.86)	-0.12 (0.92)	-0.21 (0.71)	0.22 (0.77)	-0.06 (0.88)	0.01 (0.99)	0.15 (0.80)
Punishment (victim)	-0.00 (0.84)	-0.02 (0.87)	-0.22 (0.64)	0.12 (0.82)	0.10 (1.14)	-0.07 (0.76)	0.08 (0.78)

Table 2. *Correlations among major study variables*

Variable	1	2	3	4	5	6	7	8	9	10
1. Trust Game earnings ($)										
2. Sent in Trust Game ($)	.84**									
3. Fair/accurate (insulter)	0.08	0.06								
4. Fairness/accurate (victim)	-0.11	-0.09	0.04							
5. WTR (insulter)	.26**	.32**	-0.03	.14†						
6. WTR (victim)	.38**	.39**	-0.03	.13†	.87**					
7. Anger (insulter)	-0.09	-0.08	-.20**	-.10	.05	.05				
8. Anger (victim)	-0.15	-0.14*	.19**	-.05	-.08	-.12	.25*			
9. Punishment (insulter)	0.03	-0.05	-.06	.05	-.13†	-.05	.15*	.16*		
10. Punishment (victim)	0.00	-0.03	.12†	.01	-.14†	-.16*	.36	.26**	.59*	

† = $p < .10$; * = $p < .05$; ** = $p < .01$ (not corrected for multiple comparisons).

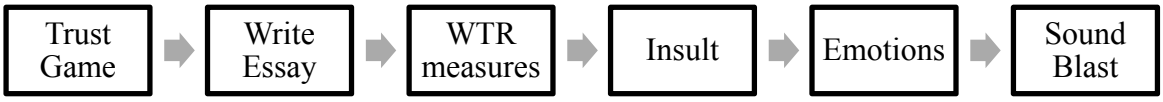

Figure 1. Timeline of experiment procedures.

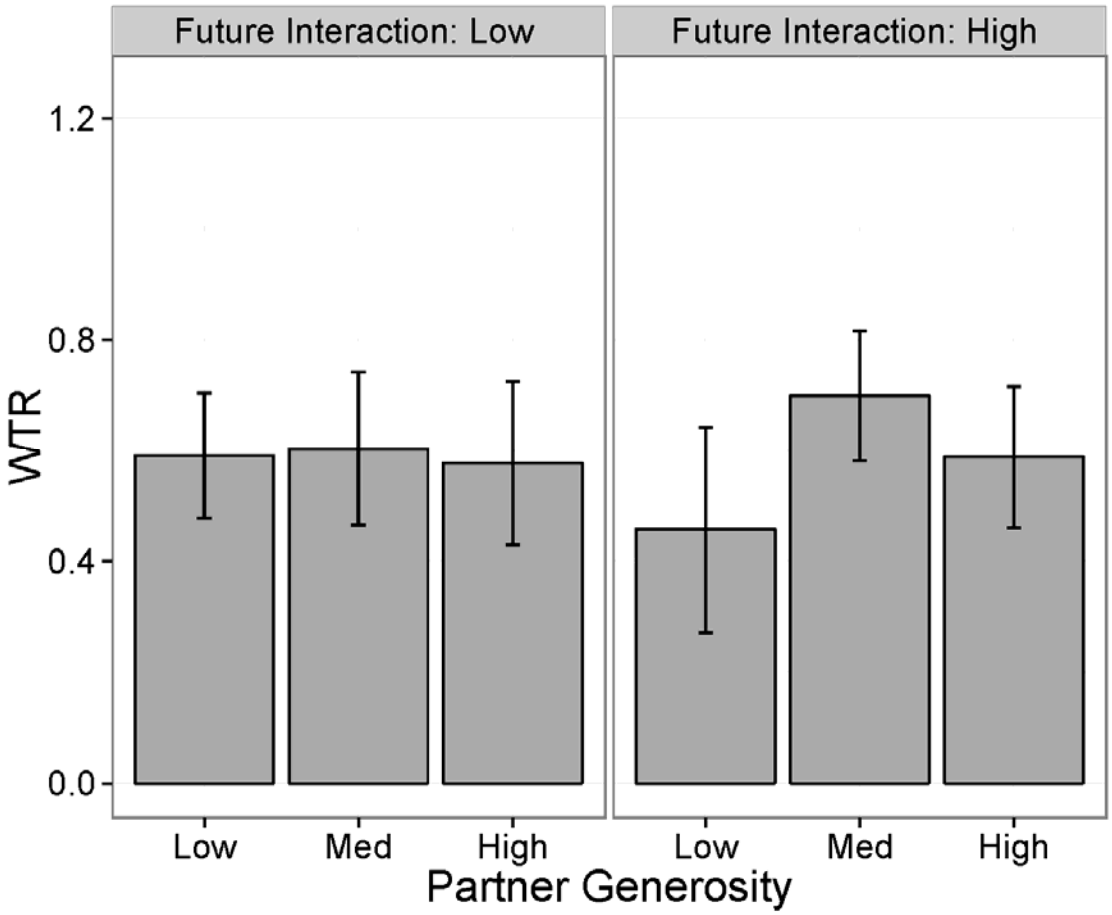

Figure 2. Mean WTR (scale: 0 to 1.13) toward the victim as a function of partner generosity and the likelihood of future interaction. Error bars indicate 95% confidence intervals.

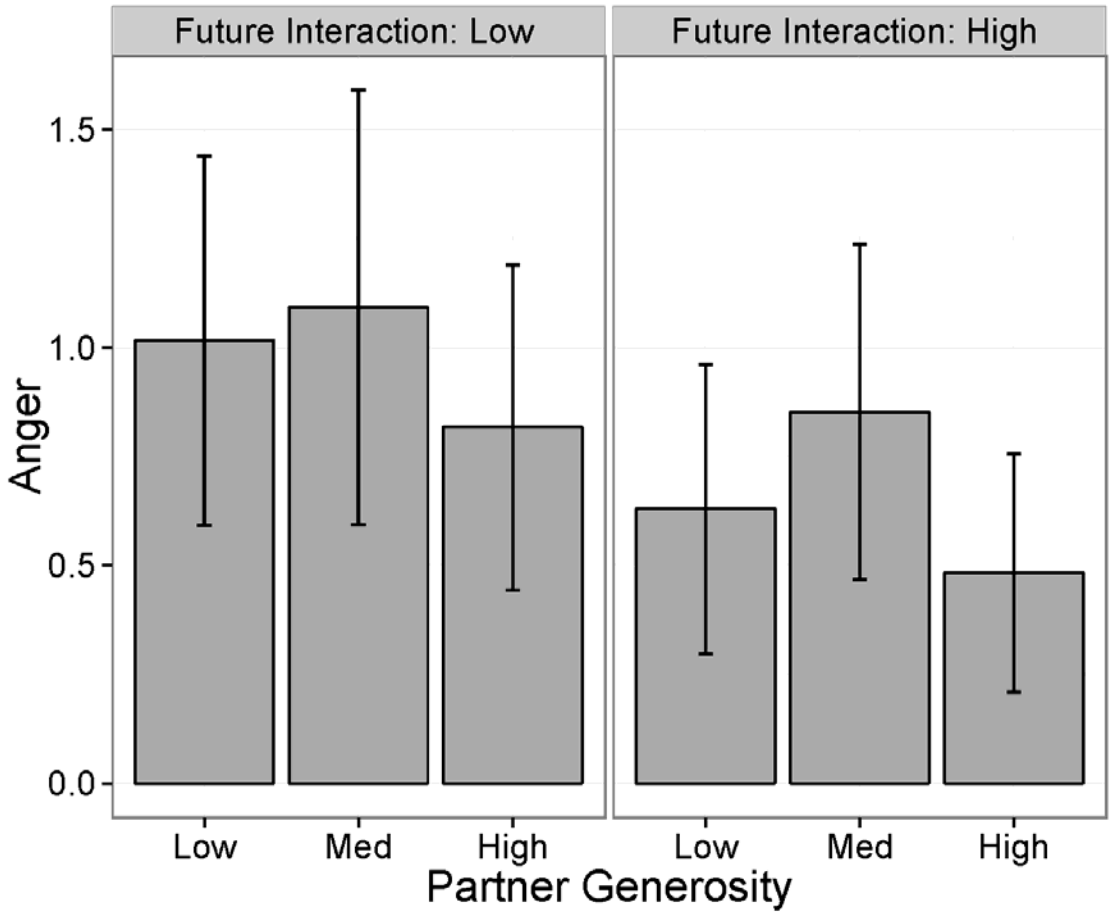

Figure 3. Mean anger (scale: 0 to 5) toward the insulter as a function of partner generosity and the likelihood of future interaction. Error bars indicate 95% confidence intervals.

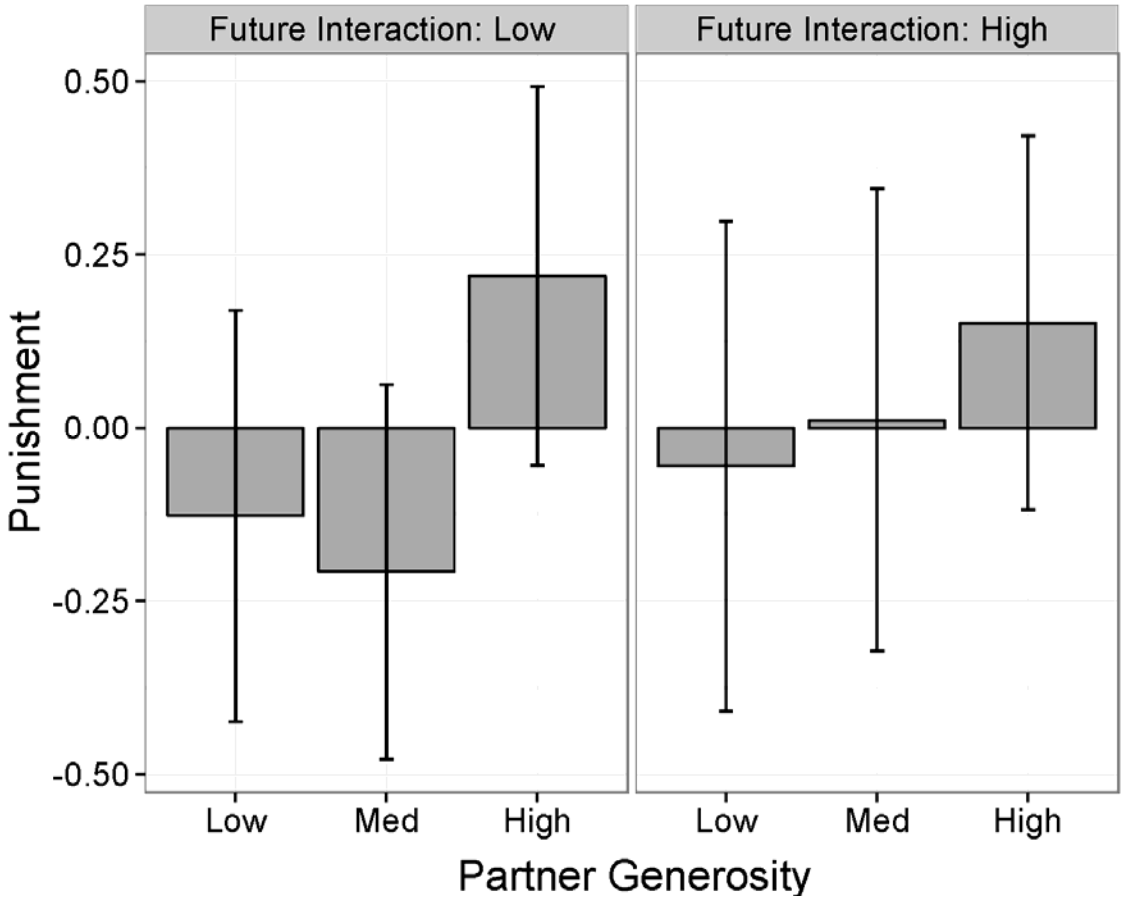

Figure 4. Mean punishment of the insulter as a function of partner generosity and the likelihood of future interaction. Since punishment is a standardized measure, a value of zero indicates punishment equal to the grand mean. Error bars indicate 95% confidence intervals.