The effects of endowment size and strategy method on third party punishment

Jillian Jordan[1*], Katherine McAuliffe[1], and David Rand[1,2,3]

[1]Psychology Department, [2]Economics Department, [3]School of Management, Yale University, New Haven CT 06511

Yale University Department of Psychology
1 Prospect Street
New Haven, CT 06511

*Corresponding author:
jillian.jordan@yale.edu

This is a pre-print of an article published in Experimental Economics. The final authenticated version is available online at: https://doi.org/10.1007/s10683-015-9466-8.

Abstract

Numerous experiments have shown that people often engage in third-party punishment (3PP) of selfish behavior. This evidence has been used to argue that people respond to selfishness with anger, and get utility from punishing those who mistreat others. Elements of the standard 3PP experimental design, however, allow alternative explanations: it has been argued that 3PP could be motivated by envy (as selfish dictators earn high payoffs), or could be influenced by the use of the strategy method (which is known to influence second-party punishment). Here we test these alternatives by varying the third party's endowment and the use of the strategy method, and measuring punishment. We find that while third parties do report more envy when they have lower endowments, neither manipulation significantly affects punishment. We also show that punishment is associated with ratings of anger but not of envy. Thus, our results suggest that 3PP is not an artifact of self-focused envy or use of the strategy method. Instead, our findings are consistent with the hypothesis that 3PP is motivated by anger.

Keywords

cooperation, norm-enforcement, strategy method, emotions, fairness, economic games

1. Introduction

Laboratory experiments using economic games have demonstrated that impartial third-party observers are often willing to pay costs to punish selfish behavior (Fehr and Fischbacher 2004; Henrich et al. 2006; Charness et al. 2008; Almenberg et al. 2010; Nikiforakis and Mitchell 2013)[1]. In these experiments, an "actor" typically has the choice to pay a cost to benefit a "recipient" (prosociality). Afterwards, a "third party" can respond to the actor's behavior by paying a cost to impose a greater cost on the actor (punishment). Many third parties choose to punish actors who behave selfishly (and selfish behavior is punished much more than fair behavior). These observations have been widely interpreted as evidence that humans not only have social preferences that lead them to act prosocially themselves, but also to intervene when others are harmed by punishing those who fail to act prosocially (Fehr and Fischbacher 2004; Jordan et al. 2014; McAuliffe et al. 2015; Henrich et al. 2006; Nikiforakis and Mitchell 2013). Furthermore, this impartial sanctioning behavior has been argued to play an important role in stabilizing human cooperation by deterring selfishness (Charness et al. 2008; Balafoutas et al. 2014; Fehr and Fischbacher 2004; Falk et al. 2005).

[1] Evidence of verbal third-party intervention in the field comes from Balafoutas and Nikiforakis 2012.

However, two elements of the standard third-party punishment (3PP) experimental design (described below) lead to potential problems in interpreting observed 3PP as evidence of displeasure over others being treated unfairly. First, in typical experiments, third parties receive small starting endowments, such that selfish actors not only out-earn second parties (whose payoffs they directly affect), but also receive higher payoffs than third-party punishers.[2] Third parties might thus be using punishment to reduce their *own* payoff disadvantage relative to actors (as punishment is more costly for the punished than the punisher), rather than to respond to the actor's treatment of the recipient (or to inequity between the actor and the recipient). If so, 3PP in these experiments would demonstrate self-focused envy rather than concern with others failing to act prosocially (Fehr and Fischbacher 2004; Pedersen et al. 2013).

Second, previous 3PP experiments may have incorrectly estimated actual willingness to punishment through their the use of the "strategy method" (Selten 1965; Brandts and Charness 2011). Under the strategy method, subjects are asked to make decisions about how to react to each possible action of the other players, prior to learning what the actions the other players actually took. In the context of 3PP, instead of responding to a specific actor behavior, punishers indicate a strategy for how much to punish each possible actor behavior. This strategy then gets implemented after the actor makes a decision.

The strategy method is popular for 3PP experiments because it reveals how each individual would respond to the full range of actor behaviors (even behaviors which actors rarely choose). However, people may behave differently in strategy method experiments than when

[2] For example, in the canonical 3PP study, participants played a dictator game with 3PP (Fehr and Fischbacher 2004). Actors received 10 monetary units, and could give up to half to the recipient. Then, third parties received 5 units to spend punishing the actor. Thus, selfish actors who gave less than half made more than 5 units, out-earning both the recipient *and* the third party.

responding to actual selfish behavior (Fischbacher et al. 2012). For example, they may under- or over-estimate how angry they would actually feel in response to selfish behavior[3], leading to an incorrect measure of actual willingness to punish (Pedersen et al. 2013). A recent review of economic games suggests that use of the strategy method does sometimes influence behavior (Brandts and Charness 2011); for example, in one study where the *recipient* was the punisher (i.e. a second-party punishment game), "hot" decisions elicited more punishment than decisions made using the strategy method (Falk et al. 2005).

Thus, two potential design confounds make interpretation of 3PP in previous laboratory experiments difficult. Does punishment really reflect a prosocial concern that the actor mistreated the recipient (be it based on norm enforcement (Fehr and Fischbacher 2004), types-based reciprocity (Levine 1998) or inequity aversion regarding the payoff differential between the actor and recipient (Fehr and Schmidt 1999))? Or does it instead reflect self-focused envy or strategy method prediction errors?

Here, we conduct two experiments addressing this issue. We systematically manipulate third-party endowments (and thus self-focused envy) and the strategy method (and thus the potential for prediction errors) in a 3PP game. We use these manipulations to test the hypothesis that 3PP is motivated by these factors rather than concerns regarding the actor's treatment of the recipient. We also investigate the association between self-reported emotions and third-party punishment. Previous research has found that negative emotions such as irritation, contempt (Bosman and Van Winden 2002) and anger (Fehr and Gachter 2002; Cubitt et al. 2011) are associated with second-party punishment. Evidence also suggests that third-party punishment is

[3] In psychology, this referred to as an "affective forecasting error", or an error in predicting how one will feel in the future (Gilbert and Wilson 2007).

associated with negative emotional reactions, such as moralistic anger (Nelissen and Zeelenberg 2009) and self-focused envy (Pedersen et al. 2013), some of which could reflect emotions experienced "on behalf" of second parties stemming from empathy or perspective taking. To investigate the role of these different processes, we measure third parties' own anger and envy, as well as their beliefs about recipients' anger and envy.

2. Methods: Experiment 1

In Experiment 1, we employed a binary dictator game (the actor could share equally or not at all) with 3PP. We manipulated whether the third party's endowment was equal to the actor's (avoiding an envy motivation) or half as large (creating an envy motivation). We crossed this with a manipulation of whether punishment decisions were "hot" responses to a particular actor choice (avoiding potential strategy method prediction errors) or made using the strategy method (allowing potential strategy method prediction errors). We also sought to directly assess the emotions motivating 3PP by asking how angry and envious third parties felt, and expected the recipient to feel, in response to the actor's behavior. This allowed us to investigate which emotions (anger versus envy) were associated with punishment, and if punishment was more strongly associated with punishers' own emotions, or the emotions they expected second parties to feel.

In addition to investigating the motivations for 3PP, we also investigated other players' expectations of, and responses to, 3PP. While there is considerable evidence that third parties punish, there is limited direct evidence of how the possibility of punishment affects selfish behavior (for exceptions see (Charness et al. 2008; Balafoutas et al. 2014)). Thus we also asked actors and recipients to predict how much third parties would punish, and investigated the association between anticipated 3PP and cooperative behavior.

2.1 Participants

Participants were recruited online through Amazon Mechanical Turk (MTurk), an online labor market in which workers complete short tasks for small payments (typically less than $1 for tasks that typically take less than 10 minutes) (Rand 2012). Employers use MTurk to "crowdsource" employees for jobs which are easy for humans but difficult for computers, such as transcribing hand-written task or classifying images. In recent years, MTurk has also become popular as a tool for experimental social scientists. MTurk jobs involve a baseline payment as well as the possibility of an additional bonus payment depending on performance, making them well-suited for economic game experiments (baseline payments correspond to show-up fees, and bonus payments are determined by the outcome of the game). MTurk may be particularly attractive to experimental economists due to participants' extremely high level of anonymity, as well as the ability to recruit a much more diverse range of subjects than the undergraduate students typical of laboratory studies.

There are, however, a number of potential issues with MTurk as an experimental platform. More importantly, experimenters necessarily sacrifice a great deal of control relative to the physical laboratory (participants might be distracted, engaged in multiple tasks at the same time, etc), and stakes are typically much smaller on MTurk than in the physical lab. To address these kinds of concerns, a large body of recent research has demonstrated the reliability of data collected using MTurk (Amir et al. 2012; Buhrmester et al. 2011; Horton et al. 2011; Mason and Suri 2012; Paolacci et al. 2010; Rand 2012; Rand et al. 2011; Suri and Watts 2011). In particular, economic game studies have found quantitative agreement between games played on MTurk (with stakes on the order of $1) and in the physical laboratory (with stakes 10 times as large), using the one-shot prisoner's dilemma (Horton et al. 2011), dictator game, public goods game,

ultimatum game, and trust game (Amir et al. 2012), and the repeated public goods game (Suri and Watts 2011)[4]. Thus, although MTurk studies involve less control and lower stakes, there is substantial evidence in support of the validity of data gathered using MTurk.[5]

2.2 Design

Participants were recruited to play an incentivized, one-shot, anonymous dictator game with 3PP. Participants were randomly assigned to the role of actor, recipient, or third-party. Participants received a show-up fee of 30 cents, as well as a bonus that was determined by their decisions. No deception was used.

Actors received 50 cents, and made a binary decision to give either 0 or 25 cents to the recipient. Then, third parties had the opportunity to punish actors, based on their decision. In a two-by-two design, we manipulated third-party endowment, and whether decisions were made "hot" or using the strategy method, resulting in four experimental conditions. Third parties were randomly assigned to receive 25 cents (*low endowment* condition) or 50 cents (*high endowment* condition). Thus in the low endowment condition, but not the high endowment condition, selfish actors (who kept 50 cents) earned more than third parties. Third parties could then spend up to 10 cents to punish the actor, based on the actor's decision. For every cent spent on punishment, the actor lost three cents.

Third parties randomly assigned to the *hot* condition were told whether the actor they were paired with gave 0 or 25 cents to the recipient, and then decided how much to punish. Third

---

[4] Other research has also shown that subjects on MTurk show high test-retest reliability on a range of personality measures (Buhrmester et al. 2011) and demographics (Mason and Suri 2012; Rand 2012), at levels comparable to college undergraduates.

[5] This limited sensitivity to stake size in economic game experiments is also consistent with other findings regarding varying the stakes in the physical lab (Camerer and Hogarth 1999) (however, we note that while manipulations of stake size often have limited effects on mean game play, they do often influence observed variance).

parties randomly assigned to the *strategy method* condition indicated, for each of the two possible actor decisions, how much they would like to punish the actor. They were informed that afterwards, they would be matched with an actor and one of their decisions would be implemented, based on the actor's choice.

2.3 Procedure

All participants began the experiment by reading the same set of instructions, in which the full rules of the game were explained. Neutral framing and language were used; punishment was described as "spending money to reduce Player 1's bonus." Participants were then asked four comprehension questions to ensure that they understood that transferring money to the recipient was costly for the actor and beneficial for the recipient, while punishing the actor was costly for both the third party and the actor.

Next, participants made their decisions. Actors decided between giving 0 or 25 cents to the recipient. Then, for each of these choices, they first predicted how much the third party would punish them (in cents), and then predicted how angry and envious the third party and recipient would each feel (on 1-7 Likert scales, ranging from "Not [angry/envious] at all" to "Very [angry/envious]").[6] Recipients predicted how much the third party would punish the actor.

Third parties were first reminded of their starting endowment. Then, in the *hot* condition, they were told how many cents the actor gave to the recipient. Next, third parties chose how much to punish, then on subsequent screens rated how angry and envious they felt, and how angry and envious they expected the recipient to feel. The order of anger and envy ratings was

[6] Although our emotion elicitations were necessarily unincentivized, there is a long tradition of using self-report emotion ratings in the social psychological literature and they have been shown to be reliable, and agree with peer ratings (Watson et al. 1988; Watson and Clark 1991). This method of measuring emotions has also been incorporated into experimental economics (Bosman and Van Winden 2002).

randomized. In the *strategy method* condition, third parties separately made each of these ratings for the cases in which the actor gave the recipient 0 or 25 cents.

Finally, all participants answered a questionnaire that included rating their confidence that the other participants were real, and indicating their age, gender, and level of education. After data from all participants was collected, actors recipients and third parties were matched into groups of three and payoffs were determined and paid accordingly (it is standard on MTurk for bonus payments to only be made once all work has been submitted and reviewed; this delay between completing the task and receiving one's bonus allows for the ex-post matching scheme we used to determine payoffs).

In the *strategy method* condition, after pairing players, we determined which third-party punishment decision to enact based on the actor's decision (to share or not share). In contrast, in the *hot* condition, we paired players based on the actor's decision (i.e. actors who shared were matched with third parties who decided how to punish sharing actors, while actors who did not share were matched with third parties who decided how to punish non-sharing actors).[7] No deception was used. For screenshots of the instructions and decision screens that were presented to subjects, see Appendix.

2.4 Statistical analysis

[7] Actors were recruited prior to third parties, so that the number of actors choosing to act selfishly or fairly was known prior to recruiting third parties. Accordingly, third parties were assigned to see selfish versus fair actor behavior in proportion to the actions of the actors. This allowed us to attached a correct 1-to-1 matching between actors and third parties.

We use linear regressions when predicting punishment in cents and emotion ratings on Likert scales, and logistic regressions when predicting (binary) actor decisions.[8] We use robust standard errors, and cluster standard errors on subject when we have repeated observations from the same subject (i.e. in the strategy method condition, in which subjects made punishment decisions about both selfish and fair offers). We exclude participants who did not answer all comprehension questions correctly, because it is unclear how to interpret the behaviour of non-comprehending subjects (Horton et al. 2011). However, we note that including them does not qualitatively change our results.[9]

3. Results: Experiment 1

3.1 Participants

N = 323 third parties (42% female, mean age = 31 years) participated and answered all comprehension questions correctly.[10]

3.2 Do third parties punish selfish behavior more than fair behavior?

We begin by confirming that selfish behavior elicits more punishment than fair behavior, collapsing across experimental conditions. We find that, as predicted, subjects spent more on punishment of selfish behavior (M = 2.08, SD = 3.63) than fair behavior (M = 0.21, SD = 1.18*)*

---

[8] Predicting emotion ratings using an ordered probit model produces qualitatively identical results; thus, we report linear regressions for consistency across analyses and ease of interpretation of coefficients.

[9] Overall, 61% of subjects answered all comprehension questions correctly (mean number of questions correct=3.34/4, with rates of comprehension on the four individual questions ranging from 75% to 93%). Thus, while a relatively low proportion of subjects answered all questions correctly, we note that subjects did relatively well on each individual question, and emphasize that all of our main results hold when including all subjects *and* when including only comprehenders. Furthermore, this rate of comprehension failure is typical for economic game studies run on MTurk (e.g. Rand et al. (2012)).

[10] Low endowment, strategy method condition N = 81; low endowment, hot condition N = 85; high endowment, strategy method condition N = 78; high endowment, hot condition N = 79.

(Figure 1). A regression predicting cents spent on punishment as a function of actor behavior (1 = selfish, 0 = fair) finds a significant positive effect of selfish behavior (coeff = 1.86, n = 323, p < .001; Table 1). Thus, as predicted, third parties systematically punished selfish over fair behavior. We also note that, as expected, there was relatively little punishment of fair behavior.

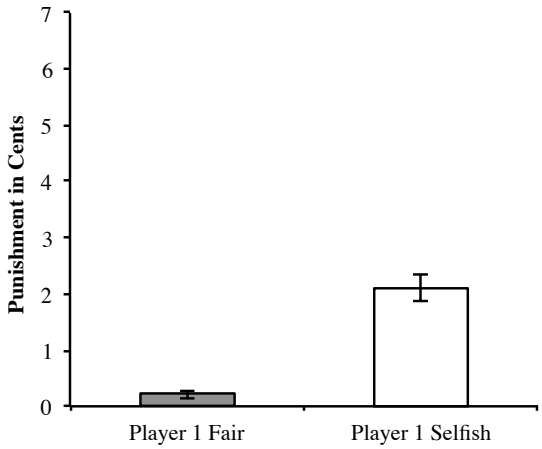


Figure 1. *Third parties respond to selfish behavior with more punishment than fair behavior in Experiment 1. Shown is the average number of cents spent by third parties on punishing fair versus selfish actor behavior, out of a maximum of 10 cents. Data collapsed across experimental conditions. Error bars indicate robust standard errors of the mean.*

	Punishment
Actor Decision (0 = Fair, 1 = Selfish)	1.863***
	(0.237)
Constant	0.214***
	(0.0788)
Observations	482
Subjects	323
Robust standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

Table 1. *This table shows the results from a linear regression predicting third-party punishment as a function of actor behavior in Experiment 1. We report the coefficients and robust standard errors clustered on subject for each independent variable. We note that there are more observations than subjects because subjects in the "strategy method" condition made two decisions, one about a selfish offer and one about a fair offer, whereas subjects in the "hot" condition made only one decision.*

3.3 Effects of endowment and strategy method manipulations

We next turn to investigating the effect of our manipulations on third-party punishment of selfishness. Because our key question is which factors led subjects to punish selfish behavior (i.e. to ask how envy and the strategy method influenced punishment of *selfish* behavior), in this analysis we focus on decisions about punishment of selfish offers[11]. In Figure 2A, we plot the mean punishment of selfishness across conditions (hot, low endowment condition: M = 2.33, SD = 3.78; cold, low endowment condition: M = 2.32, SD = 3.76; hot, high endowment condition: M = 1.63, SD = 3.58; cold, high endowment condition: M = 1.94, SD = 3.47).

We find that a regression predicting punishment of selfishness as a function of a "low endowment" dummy (1 = 25 cents, 0 = 50 cents) and a "hot" dummy (1 = hot condition, 0 = strategy method condition) finds no significant effect of the low endowment dummy (coeff = 0.509, n = 258, p = .262) or the hot dummy (coeff = -0.145, n = 258, p = .756) (Table 2 Column 1). We also find no significant interaction between the endowment and hot dummies (coeff = 0.323, n = 258, p = .730; Table 2 Column 2). Thus, our manipulations had no effect on punishment of selfishness. This suggests that punishment does not reflect (i) self-focused envy, as punishment did not increase when selfish actors earned more than third parties; or (ii) strategy method prediction errors, as punishment did not decrease when third parties made hot decisions rather than using the strategy method.

[11] Our main results are robust, however, to analyzing all decisions (i.e. punishment of both selfish and fair behavior). When including all decisions, a regression finds no significant effect of a "low endowment" dummy (coeff =0.063, n = 482, p = .813) or a "hot" dummy (coeff = 0.122, n = 482, p = .672), and a regression that adds an interaction term also finds no significant effect of the interaction (coeff = 0.241, n = 482, p = .675).

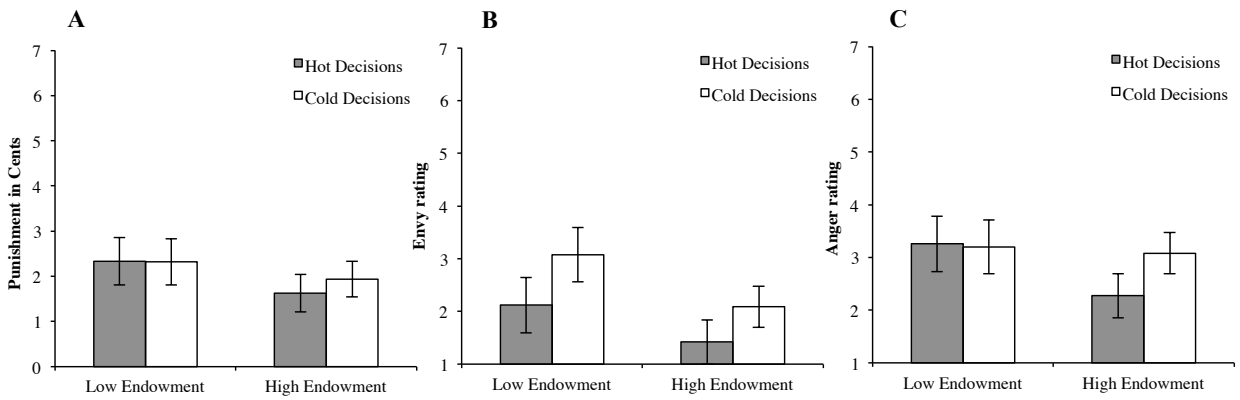


Figure 2. *The effect of our endowment and strategy method manipulations on third-party punishment, envy, and anger in Experiment 1. Shown is the mean (A) number of cents (out of a maximum of 10) spent on punishing selfish actors (B) rating of own envy, in response to selfish offers; and (C) rating of own anger, in response to selfish offers. Error bars indicate robust standard errors of the mean.*

While we found that the strategy method had no effect on punishment, one might argue that even "hot" decisions in anonymous, online experiments may not reflect the psychology of real decisions, given subjects' potential uncertainty that they were interacting with real other players. To address this concern, we asked subjects at the end of the study to rate their confidence that the other players were real (1 = very sceptical, 7 = very confident). When we repeat the above analyses including only "confident" subjects (those who reported a 5 or above, N = 95), we again find no effect of the hot dummy in a regression without an endowment interaction (coeff = -.185, n = 95, p = .819), and no hot by endowment interaction (coeff = -.439, n = 95, p = .792). Thus, even among subjects who reported being relatively confident that the other players were real, the strategy method had no effect on punishment. We also find no interaction between a hot dummy and the confidence variable when predicting punishment (coeff = 0.022, n = 258, p = .923), providing further evidence that incredulous subjects were not responsible for our finding that the strategy method had no effect on punishment.

Next, we ask how our manipulations influenced third parties' own emotional responses to selfishness. We repeat the above analyses with own envy and anger, rather than punishment, as dependent variables. Beginning with envy, in Figure 2B, we plot mean envy in response to selfishness across conditions (hot, low endowment condition: M = 2.12, SD = 1.61; cold, low endowment condition: M = 3.07, SD = 2.13; hot, high endowment condition: M = 1.42, SD = 1.07; cold, high endowment condition: M = 2.09, SD = 1.58). In regression analysis, we find a significant positive effect of the low endowment dummy (coeff = 0.876, n = 258, $p < .001$) and a significant negative effect of the hot dummy (coeff = -0.818, n = 258, $p < .001$) (Table 2 Column 3), and no significant interaction (coeff = -0.283, n = 258, p = .482; Table 2 Column 4).

Thus, our manipulations significantly influenced envy. First, participants in the low endowment condition reported more envy. Critically, this increase serves as a manipulation check, suggesting that third parties did actually attend to their endowment, and felt more envious when selfish actors earned more than them. This manipulation check confirms that our endowment manipulation successfully increased envy but did not increase punishment, suggesting that envy does not motivate punishment. Second, participants in the strategy method condition also reported more envy. This suggests that envy may in part be an artifact of the strategy method, rather than a genuine reaction to unfairness.

We next investigate anger. In Figure 2C, we plot mean anger in response to selfishness across conditions (hot, low endowment condition: M = 3.25, SD = 2.12; cold, low endowment condition: M = 3.20, SD = 2.09; hot, high endowment condition: M = 2.27, SD = 1.70; cold, high endowment condition: M = 3.08, SD = 1.90). In regression analysis, we find no significant effects of the endowment dummy (coeff = 0.452, n = 258, p = .067) or the hot dummy (coeff = -

0.363, n = 258, p = .150) (Table 2 Column 5), and no significant interaction (coeff = 0.863, n = 258, p = .084; Table 2 Column 6) between the two.

However, we note that effects of the endowment dummy, and the interaction term, are both *marginally* significant; Table 2 Column 6 demonstrates that this is driven by a significant effect of decision method within the high endowment condition (with subjects reporting less anger in the "hot" condition). Thus, anger did not vary significantly across conditions, although there was a trend in the direction of subjects reporting less anger when they had high endowments and made hot decisions. This may suggest that subjects in the high endowment condition made an affective forecasting error in which they expected to experience more anger than they actually did.

	(1)	(2)	(3)	(4)	(5)	(6)
	Punishment	Punishment	Envy	Envy	Anger	Anger
Endowment size (0 = High, 1 = Low)	0.509	0.385	0.876***	0.984***	0.452*	0.121
	(0.453)	(0.574)	(0.21)	(0.297)	(0.246)	(0.317)
Decision method (0 = Strategy method, 1 = Hot)	-0.145	-0.311	-0.818***	-0.673***	-0.363	-0.806**
	(0.467)	(0.649)	(0.203)	(0.236)	(0.251)	(0.326)
Endowment size X Decision method		0.323		-0.283		0.863*
		(0.935)		(0.403)		(0.498)
Constant	1.873***	1.936***	2.145***	2.090***	2.908***	3.077***
	(0.357)	(0.393)	(0.16)	(0.179)	(0.195)	(0.215)
Observations	258	258	258	258	258	258
Subjects	258	258	258	258	258	258
Robust standard errors in parentheses						
*** p<0.01, ** p<0.05, * p<0.1						

Table 2. *This table shows the results from linear regressions predicting third-party punishment (Columns 1-2), envy (Columns 3-4), and anger (Columns 5-6) in response to selfish actor behavior, as a function of endowment size and decision method in Experiment 1. We report the coefficients and robust standard errors clustered on subject for each independent variable.*

3.4 Which emotions predict individual differences in third-party punishment?

We now directly ask which emotions were associated with punishment by examining the relationship between individual emotion ratings and punishment of selfishness. In this analysis, we consider punishment of either selfish or fair offers as our dependent variable. We analyze punishment of all offers because we hypothesize that the *reason* that selfish offers were punished more than fair offers is that they elicited more negative emotional reactions; thus, it makes sense to consider the variance in emotional reactions, and punishment, across all offers. We conduct a regression predicting punishment as a function of a low endowment dummy, a hot dummy, the third party's own anger and envy, and the anger and envy the third party predicted that the recipient would experience. We find that third-party punishment shows a significant positive association with own anger (coeff = .807, n = 323, $p < .001$), a significant negative association with own envy (coeff = -0.264, n = 323, p = .006), and no significant association with predicted recipient anger (coeff = .138, n = 323, p = .376) or envy (coeff = -0.025, n = 323, p = .863) (Table 3 Column 1).

Thus, across experimental conditions and actor transfers, only one emotion variable was positively associated with punishment: Participants who reported *themselves* being angrier spent more on punishment, while there was no significant positive association with envy or attributed recipient emotions. We also note that own anger continued to be significantly associated with punishment when considering each condition separately (Table 3 Columns 2-5).

Interestingly, participants who reported stronger feelings of envy actually spent *less* on punishment, when controlling for their own anger and predicted recipient emotions. This effect was unexpected, and partitioning data by experimental condition reveals that it is driven by the [strategy method, low endowment] condition. In this condition, there is a strong negative

association between punishment and envy (coeff = -0.564, n = 81, $p < .001$) (Table 3 Column 2), while the other three conditions reveal no significant associations (all p values > .3) (Table 3 Columns 3-5). We return to this apparent negative association with envy in Experiment 2.

Figure 3 shows the association between one's own envy and anger and punishment. To visualize the independent associations with each variable, we perform a median split on own anger and own envy, and divide participants into four groups accordingly. Figure 3 shows that participants reporting above-median anger punished much more than participants reporting below-median anger, regardless of envy levels. In contrast, participants reporting above-median envy spent slightly *less* than participants reporting below-median envy, regardless of anger levels. In sum, then, our analyses suggest that 3PP is associated with the third party's level of anger, and not their level of envy.

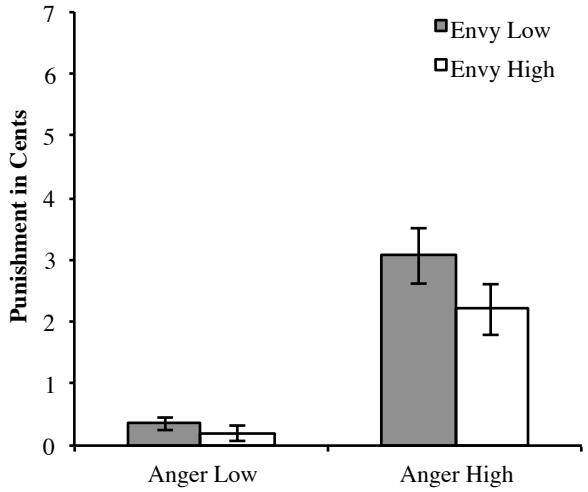


Figure 3. *High anger, but not high envy, is associated with third-party punishment of selfishness in Experiment 1. Shown is the average number of cents (out of a maximum of 10) third parties spent on punishing selfish actors, by their anger and envy ratings. For ease of visualization, median splits on emotional ratings are shown. Data collapsed across experimental conditions. Error bars indicate robust standard errors of the mean.*

	(1)	(2)	(3)	(4)	(5)
	Punishment	Punishment	Punishment	Punishment	Punishment
		(Low endowment, Strategy method)	(Low endowment, Hot)	(High endowment, Strategy method)	(High endowment, Hot)
Player 3 Anger	0.807***	0.768***	0.911***	0.910***	0.805**
	(0.124)	(0.179)	(0.272)	(0.222)	(0.334)
Player 3 Envy	-0.264***	-0.564***	0.0107	-0.169	-0.157
	(0.0952)	(0.137)	(0.247)	(0.187)	(0.268)
Predicted Player 2 Anger	0.138	-0.000579	0.0541	-0.141	0.539*
	(0.156)	(0.306)	(0.315)	(0.183)	(0.273)
Predicted Player 2 Envy	-0.0249	0.325	-0.0267	0.126	-0.536**
	(0.144)	(0.281)	(0.313)	(0.170)	(0.261)
Endowment size (0 = High, 1 = Low)	0.0232				
	(0.235)				
Decision method (0 = Strategy method, 1 = Hot)	-0.0319				
	(0.237)				
Constant	-0.418*	-0.362	-0.929**	-0.454	-0.122
	(0.241)	(0.239)	(0.435)	(0.417)	(0.776)
Observations	482	162	85	156	79
Subjects	323	81	85	78	79
Robust standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					

Table 3. *This table shows the results from linear regressions predicting third-party punishment as a function of third-party emotions, and predicted second-party emotions in Experiment 1. We report the results collapsed across conditions (Column 1) as well as separately by condition (Columns 2-5). We report the coefficients and robust standard errors clustered on subject for each independent variable.*

3.5 Actor and recipient responses

Finally, we analyze the responses of actors and recipients. We find that both actors and recipients expected third parties to punish selfishness more than fairness. Interestingly, actors and recipients in fact significantly *over*-estimated how much third parties would punish

selfishness. We also find that actors who anticipated more 3PP of selfishness, relative to fairness, were less likely to be selfish (presumably in order to avoid getting punished). These results suggest that people anticipate 3PP, and that anticipated punishment may motivate fair behavior. For a more detailed discussion of these results, see Appendix.

4. Discussion: Experiment 1

Experiment 1 suggests that third-party punishment is not an artifact of self-focused envy or the strategy method. We found that third-party punishment was not influenced by manipulating third-party endowments, despite the fact that third parties with low endowments reported more envy than third parties with high endowments. Third-party punishment was also not influenced by manipulating the use of the strategy method, in contrast to evidence that the strategy method reduces levels of second-party punishment (Falk et al. 2005). Furthermore, anger, but not envy, was associated with individual differences in punishment: individual subjects who reported experiencing more anger also punished more. Interestingly, we found that subjects' *own* anger ratings, rather than their predictions of *recipients'* negative emotions, were what tracked punishment. Together, these results suggest that third parties experience anger when others are harmed, and that their own anger is associated with their decisions to engage in third-party punishment. We also provide evidence that others anticipate such punishment, even more than it actually occurs, and that anticipated punishment is associated with fair actor behavior.

These results leave three important open questions. First, while we interpreted the finding that low third-party endowments did not increase 3PP as evidence that punishment was not motivated by envy, an alternative explanation is possible: while third parties in the low endowment condition had a stronger envy motivation (because they earned less than selfish

actors), they also had a smaller income to spend on punishment. If having a low endowment makes third parties more envious (*increasing* punishment) but also less willing to spend their (smaller) income on punishment (*decreasing* punishment), these two effects could cancel each other to result in no net effect of our endowment manipulation (as we observed). Thus, it is not clear if such an income effect confounded our results. Second, we did not predict that envy would negatively predict 3PP, and it is not clear how robust this effect is. Finally, many 3PP experiments allow actors to choose between a range of relatively fair and relatively selfish behaviors (Fehr and Fischbacher 2004; Henrich et al. 2006; Bernhard et al. 2006), while actors in our experiment made only a binary decision to share nothing or half. Thus, it is not clear if our results would replicate in a game where actors face a continuous range of decisions about how much to share with recipients.

5. Method: Experiment 2

In Experiment 2, we addressed these questions. In addition to asking whether the unanticipated negative association between envy and punishment observed in Experiment 1 would replicate, we made two changes to the experimental design. First, we added an additional condition to ask whether punishment would increase if we doubled the endowments that actors and third parties began with. This condition thus allowed us to investigate whether the null result of our endowment manipulation in Experiment 1 resulted because subjects in the low endowment condition were disinclined to spend their (smaller) income on punishment, counteracting an effect of envy.

Second, we allowed actors to decide how much money to transfer, in 10-cent increments, and tested whether our results from Experiment 1 would replicate. Because running a "hot" experiment with a large set of actor choices would require a very large sample, and because

Experiment 1 revealed that the strategy method did not influence punishment, we eliminated the "hot" condition in Experiment 2.

Thus, Experiment 2 was identical to Experiment 1, with the following exceptions. First, we had three experiment conditions. In the *high-high condition,* both the third party and actor received high endowments: they each started with 100 cents, and there was thus no envy motivation for 3PP. In the *low-low* condition, both the third party and actor received low endowments: they each started with 50 cents, and thus there was again no envy motivation for 3PP. However, because endowments were half as large, a comparison between these conditions allowed us to investigate whether third parties punish less when they have lower endowments. Finally, in the *low-high condition,* the third party received a low endowment while the actor received a high endowment: the third party started with 50 cents, while the actor started with 100 cents. Thus, selfish actors (who kept more than half) earned more than third parties, providing an envy motivation for punishment.

Second, all third parties made their decisions using the strategy method. For each of the six possible actor transfers, third parties first indicated how much to punish, and then indicated, in a random order, how angry and envious they would feel. For simplicity, we did not ask third parties how angry and envious they expected recipients to feel, as we found no significant effects of these ratings in Experiment 1. We note that due to a technical error, emotion ratings were collected incorrectly for subjects in the "low-low" condition and were thus not analyzed. We analyzed our data using the same approach as in Experiment 1, again restricting to comprehending subjects, and using linear regressions with robust, clustered standard errors.[12]

[12] We note that as in Experiment 1, our analyses predicting emotion ratings produce qualitatively equivalent results using ordered probit regressions; we thus again report only linear regression.

419

420

421 **6. Results and Discussion: Experiment 2**

422 N = 153 third parties (24% female, mean age = 28 years) recruited using Amazon
423 Mechanical Turk answered all comprehension questions correctly.[13]

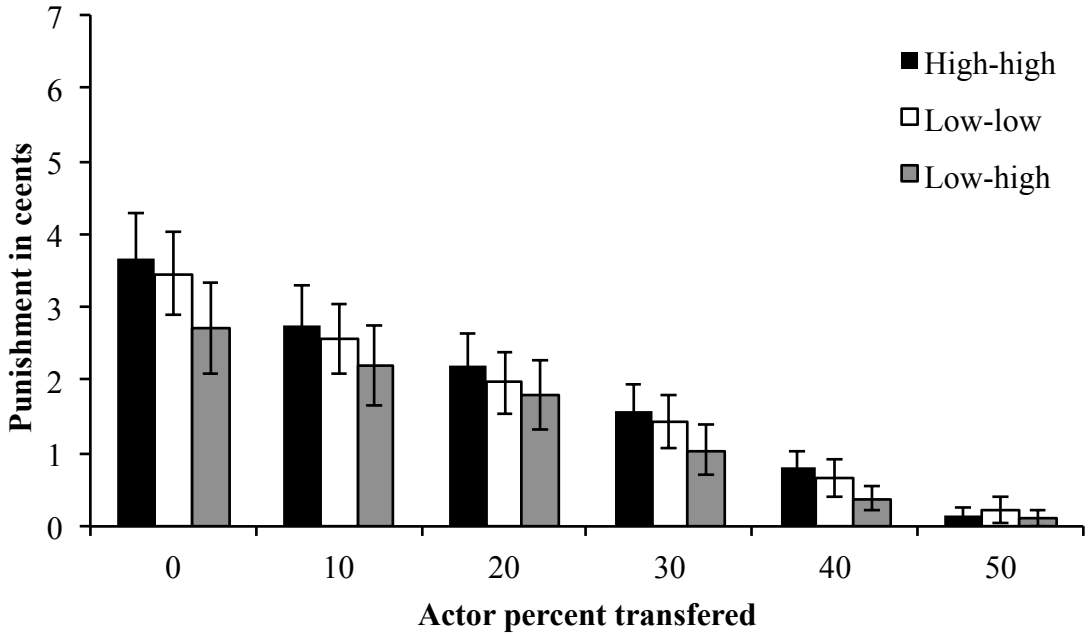
424 We begin by replicating the finding that third parties respond to unfair behavior with
425 more punishment than fair behavior. A regression predicting punishment as a function of cents
426 transferred by the actor reveals a significant negative effect of cents transferred (coeff = -0.629, n
427 = 153, $p < .001$), suggesting that selfish transfers were punished more harshly.

428 Next, we investigate the effects of our endowment manipulation on punishment, anger,
429 and envy. Because we no longer have a clear binary separation between "selfish" and "fair" actor
430 transfers, we analyze all decisions (i.e. responses to all actor transfers). In each regression, we
431 use actor transfer, a condition dummy, and the interaction between these two as independent
432 variables. In these analyses, the condition dummy term indicates the effect of condition on
433 punishment of the most selfish behavior (transferring 0 cents), and the interaction term indicates
434 whether the effect of condition changes as a function of the actor's transfer.

435 We begin by investigating punishment. We plot mean punishment across conditions, for
436 each actor transfer, in Figure 4 (punishment in response to maximum selfishness: high-high
437 condition: M=3.67, SD=4.53; low-low condition: M=3.46, SD=4.36; low-high condition:
438 M=2.70, SD=4.20). We first investigate the effect of envy on punishment by comparing our two
439 "no envy" conditions (high-high and low-low) to our "envy" condition (low-high). We find no
440 significant effect of the envy condition dummy (coeff = -0.699, n = 153, p = .371) or interaction

[13] High-high condition N = 52; low-low condition N = 57; low-high condition N = 44.

441 between actor transfer and envy condition dummy (coeff = .115, n = 153, p = .455) (Table 4
442 Column 1).



443

444 **Figure 4.** *Subjects punish equally across our three endowment in Experiment 2. Shown is the*
445 *average number of cents (out of a maximum of 10) spent by third parties on punishing actors, by*
446 *endowment condition and actor transfer. Error bars indicate robust standard errors of the mean.*
447

448 Next, we ask whether this result holds when comparing our "envy" (low-high) condition
449 to both of the "no envy" (high-high and low-low) conditions separately, and find that it does
450 (comparison to high-high condition: no effect of the envy dummy (coeff = -0.838, n = 96, p =
451 .360) or interaction (coeff = .142, n = 96, p = .431); comparison to low-low condition: no effect
452 of the envy dummy (coeff = -0.572, n = 101, p = .513) or interaction (coeff = .091, n = 101, p =
453 .596). Thus, we replicate our finding from Experiment 1 that third parties do not punish
454 significantly more when envy motivations are possible.
455 Finally, we compare the high-high condition to the low-low condition to investigate a
456 possible income effect on punishment. We find no significant effect of a high-high dummy

(coeff = .266, n = 109, p = .757) or the interaction between a high-high dummy and actor transfer (coeff = -.050, n = 109, p = .761) on punishment (Table 4 Column 2). Thus, third-party punishment does not appear to be sensitive to income (at least over the range of values we consider here): doubling endowments had no effect on punishment. This suggests that Experiment 1 was not confounded by an income effect.

	(1)	(2)
	Envy contrast	Income contrast
Actor transfer	-0.663***	-0.790*
	(0.082)	(0.430)
Endowment condition	-0.699	0.266
	(0.779)	(0.858)
Actor transfer X Endowment condition	0.115	-0.050
	(0.154)	(0.166)
Constant	3.443***	4.11*
	(0.427)	(2.23)
Observations	918	654
Subjects	153	109
Robust standard errors in parentheses		
*** p<0.01, ** p<0.05, * p<0.1		

Table 4. *This table shows the results from linear regressions investigating the effects of envy (Column 1) and income (Column 2) on punishment. Both regressions predict third-party punishment as a function of endowment condition, actor transfer, and their interaction in Experiment 2. Column 1 shows the effect of envy (0 = Envy not possible (high-high and low-low conditions), 1=Envy possible (low-high condition)); Column 2 shows the effect of income (0 = low income (low-low condition), 1 = high income (high-high condition)). We report the coefficients and robust standard errors clustered on subject for each independent variable.*

We next turn to investigating the effects of our endowment manipulation on emotion ratings; we again note that these analyses exclude the "low-low" condition where emotions were incorrectly measured due to a technical error. We first investigate the effect on envy ratings. In regression analysis, we find a significant positive effect of a low-high condition dummy (coeff =

1.67, n = 96, p = .001), indicating more envy in this condition when actors transfer nothing (high-high condition, M=2.44, SD=2.15; low-high condition, M=3.98, SD=2.50), and a significant negative interaction between the low-high condition dummy and actor transfer (coeff = -0.263, n = 96, p = .009), indicating that this effect is stronger when the actor transfers less (and thus earns relatively more than the third party) (Table 5 Column 1). This again serves as a manipulation check, demonstrating that third parties compared their payoffs to actors, and felt envious when they had relatively less.

Next, we investigate the effects of our manipulation on anger ratings. In regression analysis, we find no significant endowment effect (coeff = .209, n = 96, p = .682; anger when actor transfers nothing: high-high condition, M=3.85, SD=2.24; low-high condition, M=4.05, SD=2.31) or interaction (coeff = -0.033, n = 96, p = .743) (Table 5 Column 2) when predicting anger. Thus, replicating Experiment 1, we find that anger is not significantly influenced by third-party endowment size, whereas envy is.

	(1)	(2)
	Envy	Anger
Actor transfer	-0.192***	-0.566***
	(0.0628)	(0.0626)
Endowment condition (0 = High-high, 1 = Low-high)	1.931***	0.243
	(0.577)	(0.606)
Actor transfer X Endowment condition	-0.263***	-0.0334
	(0.0985)	(0.102)
Constant	2.553***	4.526***
	(0.351)	(0.400)
Observations	576	576
Subjects	96	96
Robust standard errors in parentheses		
*** p<0.01, ** p<0.05, * p<0.1		

Table 5. *This table shows the results from linear regressions predicting envy (Column 1) and anger (Column 2) as a function of endowment condition, actor transfer, and their interaction in Experiment 2. We report the coefficients and robust standard errors clustered on subject for each independent variable. We again note that this analysis excludes the "low-low" condition, in which a technical error influenced the recording of emotions.*

Finally, we replicate the finding that elevated anger, but not envy, is associated with punishment. We regress punishment (of any offer) in the high-high and low-high conditions (in which emotion data was reordered correctly) against a low-high endowment dummy and anger and envy ratings. We find a significant positive association with anger (coeff = .890, n = 96, p < .001) and no significant association with envy (coeff = -.032, n = 96, p = .810) (Table 5 Column 1). We find similar results considering each experimental condition separately (Table 5 Columns 2-3). Thus, we replicate the effect that anger is associated with punishment. We do not, however, replicate the unanticipated finding from Experiment 1 that envy was negatively associated with punishment. Thus we conclude that this latter finding was likely spurious.

	(1)	(2)	(3)
	Punishment	Punishment	Punishment
		(Low-high)	(High-high)
Anger	0.890***	1.002***	0.803***
	(0.145)	(0.217)	(0.191)
Envy	-0.0321	-0.246	0.251
	(0.133)	(0.167)	(0.216)
Endowment condition (0 = High-high, 1 = Low-high)	-0.563		
	(0.424)		
Constant	-0.345	-0.590*	-0.658
	(0.385)	(0.317)	(0.488)
Observations	576	264	312
Subjects	96	44	52
Robust standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

Table 6. *This table shows the results from linear regressions predicting third-party punishment as a function of third-party emotions in Experiment 2. We report the results collapsed across conditions (Column 1) as well as separately by condition (Columns 2-3). We report the coefficients and robust standard errors clustered on subject for each independent variable. We again note that this analysis excludes the "low-low" condition, in which a technical error influenced the recording of emotions.*

7. General Discussion

Third parties punish selfish behavior in laboratory experiments, but possible design confounds have left open the question of whether this punishment reflects a true distaste for unfair treatment of third parties. Here, we provide evidence suggesting that 3PP is *not* an artifact of self-focused envy or the use of the strategy method, and may in fact reflect genuine anger that recipients were treated selfishly. Across two experiments, we support this conclusion through two main findings. First, third parties responded to selfish behavior with as much punishment and anger when their endowments were equal to actors' endowments (ruling out envy motivations) and when they made "hot" decisions (ruling out strategy method prediction

errors).[14] Second, individual ratings of one's own anger, but not envy, were associated with individual levels of punishment.

Our results have important implications for the role of punishment in promoting cooperative behavior: they are consistent with the hypothesis that impartial third-party observers react to selfishness with anger that motivates 3PP. This would suggest that third parties may indeed incur costs to punish selfishness in a variety of real-world contexts, even when they have not been directly disadvantaged. Our experiments do not, however, distinguish between different "prosocial" motivations for 3PP. For example, the anger and punishment we observe might be caused by displeasure over norms being violated (Fehr and Fischbacher 2004), by motives stemming from types-based reciprocity (Levine 1998) whereby people get utility from harming "bad" people, or by displeasure over the inequity that exists between selfish actors and their recipients (Fehr and Schmidt 1999). Distinguishing between these possibilities is an important direction for future work.

Our results build on previous research concerning the influence of possible design confounds in 3PP experiments. While most 3PP experiments have employed low third-party

[14] One might argue that it is difficult to draw strong inferences from the finding that our manipulations of endowment and the strategy method did *not* influence punishment, because they were null results. However, we note that we replicated the null finding that endowments did not influence punishment in both Experiment 1 and 2. Furthermore, our endowment manipulation *did* have a significant positive effect on envy ratings, providing a positive control that demonstrates that subjects were sensitive to the manipulation. We also conduct a power analysis to assess the smallest effects of our endowment and strategy method manipulations that we could have detected with 80% probability in Experiment 1. We find that smallest detectable effects are (i) a 1.27-cent decrease in punishment in the high endowment relative to the low endowment condition, and (ii) a 1.32-cent decrease in punishment in the strategy method condition relative to the "hot" condition. Thus, while it is possible that we failed to detect a true but small effect of these variables on punishment, this analysis provides a likely upper bound for the size of these effects, and suggests that the use of low endowments or the strategy method cannot fully account for punishment in these conditions.

endowments, such that selfish actors earned more than third parties (Fehr and Fischbacher 2004; Henrich et al. 2006; Henrich et al. 2010; Marlowe et al. 2008; Bernhard et al. 2006; Nelissen and Zeelenberg 2009; Almenberg et al. 2010; Shinada et al. 2004; Kurzban et al. 2007), others have avoided this possible confound and still observed punishment of selfish behavior (Götte et al. 2006; Bruene et al. 2012; Fehr and Fischbacher 2004; Balafoutas et al. 2014)**.** Additionally, one study directly manipulated third-party endowment, but found no significant non-zero punishment in either endowment condition (perhaps because it employed a non-standard design), leaving open the question of what motivates punishment when it is observed (Pedersen et al. 2013). Here, we provide the first direct test of this question by using the standard method but varying endowment, and find no evidence that envy motivates punishment.

With respect to strategy method prediction errors, while many 3PP experiments have employed the strategy method (Bernhard et al. 2006; Fehr and Fischbacher 2004; Henrich et al. 2010; Marlowe et al. 2008; Almenberg et al. 2010; Henrich et al. 2006), others have not (Nelissen and Zeelenberg 2009; Shinada et al. 2004; Kurzban et al. 2007) and still observed punishment of selfishness. One study of second-party punishment found that the strategy method *decreased* punishment (Falk et al. 2005); conversely, another study found that, consistent with strategy method prediction errors, participants who read a hypothetical description of a 3PP game reported that they would respond to selfishness with more anger and punishment than real third parties actually did in a different lab experiment (Pedersen et al. 2013).

Here, we provide the first direct manipulation of the strategy method in an incentivized, non-hypothetical 3PP experiment. We find no evidence that the strategy method influences punishment. Thus, our results differ from Falk and colleagues' (2005) 2PP experiment, perhaps suggesting that 2PP is driven by different motivators than 3PP (Crockett et al. 2013). Our results

559 also differ from Pedersen and colleagues' (2013) hypothetical experiment, suggesting that
560 incentivized decisions through the strategy method are not equivalent to decisions in a
561 hypothetical game.

562 Our results also build on previous work investigating emotions in 3PP experiments. In
563 some previous research, third parties have responded to selfish behavior with anger and
564 punishment (Nelissen and Zeelenberg 2009), and in others, third parties have responded with
565 envy, but not anger or punishment (Pedersen et al. 2013). These results are consistent with the
566 hypothesis that anger but not envy is necessary to motivate punishment, but leave open the
567 question of why selfishness elicits different emotional responses in different experiments.
568 Differences may result from variation in experimental designs (for example, in (Pedersen et al.
569 2013), actor behavior and 3PP behavior took place in separate interactions, and emotions were
570 assessed before punishment decisions) or subject pools.

571 Finally, our results also provide direct evidence about the 'pacifying' effect of 3PP on
572 potential selfish actors. For punishment to deter selfish behavior, individuals must perceive a
573 strong threat of punishment. Indeed, we found that actors and recipients expected third-party
574 observers to punish selfish behavior, even more harshly than they actually did, and that actors
575 who anticipated more punishment cooperated more. Furthermore, although the average amounts
576 of observed 3PP were fairly low in both experiments, many individual punishers punished the
577 maximum amount allowed (44% of punishers in Experiment 1, 63% in Experiment 2). This
578 provides additional support for the hypothesis that 3PP may discourage selfish behavior in the
579 real world. However, we note that the observed association between cooperation and expected
580 punishment was correlational, and does not establish causality. Using manipulation studies to
581 build on these results is an important direction for future research.

582 Likewise, while our results demonstrate that self-reported anger is associated with third-
583 party punishment, they leave open the question of whether anger actually *causes* punishment.
584 While our results are consistent with the hypothesis that anger causes punishment, it is also
585 possible that punishing makes subjects angry, or that unmeasured third variables (e.g. other
586 unmeasured emotions, such as empathy for the recipient, or disappointment towards the dictator)
587 cause subjects to experience anger *and* engage in punishment. Alternatively, subjects may have
588 reported feeling anger without actually having experienced it (for e.g., if subjects believe,
589 explicitly or implicitly, that anger is a socially desirable motivation to punish). To address these
590 possibilities, future research should investigate the causal role of anger on punishment by
591 inducing (or attenuating) anger before giving subjects the opportunity to engage in 3PP.
592 Furthermore, if anger appears to cause 3PP, future studies should investigate the processes by
593 which anger arises in response to selfish behavior.
594 We also acknowledge that our results reflect play in anonymous experiments on Amazon
595 Turk, with relatively low stakes. Future research should investigate if envy may influence
596 punishment in situations that are more naturalistic, or in which the stakes are higher (and thus the
597 payoff differences between selfish actors and third parties are higher). While there is substantial
598 evidence that economic game play on Mturk is largely consistent with play in the physical
599 laboratory (see introduction), it is possible that the effect of envy on punishment behavior is
600 dependent on stakes, or would be larger in a less anonymous or more naturalistic context.
601 In conclusion, 3PP of selfish behavior is frequently observed in laboratory experiments
602 and is cited as evidence that people dislike it when others fail to act prosocially, even when they
603 themselves are not harmed as a consequence. Here, we support this interpretation by providing

evidence that 3PP is not an artifact of self-focused envy or the strategy method, and may reflect genuine anger caused by selfish actions.

8. Appendix

8.1 Extended analyses of actor and recipient behavior in experiment one

Here, we report more detailed analyses of actor and recipient behavior in experiment one. N = 269 actors (45% female, mean age = 32 years) and N = 300 recipients (42% female, mean age = 30 years) participated and answered all comprehension questions correctly.

First, we ask if other players expect third parties to punish, and how anticipated 3PP compares to actual 3PP. Figure 5 plots anticipated and actual punishment across players for fair and selfish actor transfers, and demonstrates that both actors and recipients expect third parties to punish, and to punish more for selfish than fair decisions. Indeed, regressions predicting punishment as a function of selfish behavior demonstrates that both actors (coeff = 3.64, n = 269, $p < .001$) and recipients (coeff = 2.55, n = 300, $p < .001$) expect third parties to punish selfishness more than fairness.

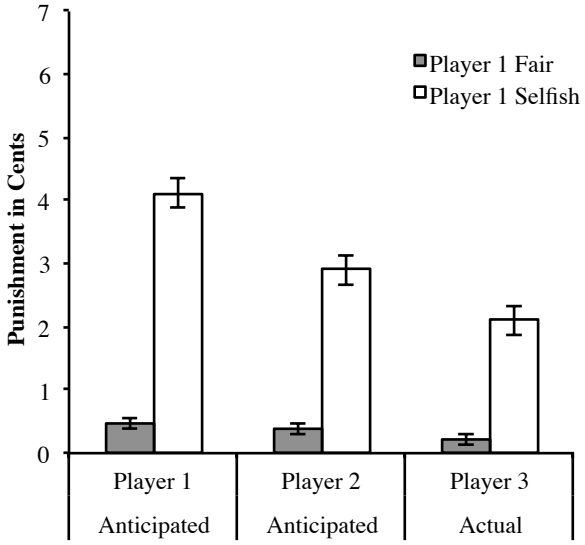


Figure 5. *Actors and recipients anticipate more third-party punishment than actually occurs in Experiment 1. Shown is the average number of cents (out of a maximum of 10) that actors and recipients anticipated third parties would spend on punishing selfish and fair actors, as well as the amount actually spent by third parties. Data collapsed across all experimental conditions. Error bars indicate robust standard errors of the mean.*

Next, we ask how anticipated punishment compares to actual punishment. We conduct a regression predicting punishment of selfishness as a function of player dummies. We find that compared to third parties as the baseline, there is a significant positive effect of an actor dummy (coeff = 2.03, n = 758, $p < .001$) and of a recipient dummy (coeff = 0.823, n = 758, $p = .015$), indicating that both actors and recipients anticipate more punishment of selfishness than actually occurs. Thus, actors and recipients anticipate high levels of 3PP targeted at selfishness.

Finally, we ask if anticipated 3PP motivates actors to share with recipients. For each individual actor, we calculate the difference between expected 3PP for selfish and fair behavior. We also calculate the difference between expected second- and third-party emotional responses to selfish and fair behavior. We then conduct a regression predicting the actor's decision to share

638 with the recipient as a function of these variables, controlling for third-party endowment. (We do
639 not control for strategy method condition, as this manipulation did not apply to actors.)
640 We find a significant positive effect of differences in expected 3PP (coeff = .173, n =
641 269, $p < .001$) and third-party anger (coeff = .281, n = 269, p = .005), a significant negative
642 effect of differences in expected second-party anger (coeff = -.245, n = 269, p = .018), and no
643 significant effect of differences in expected second-party envy (coeff = -.063, n = 269, p = .373)
644 or third-party envy (coeff = .070, n = 269, p = .250). Thus, anticipated 3PP and anger appear to
645 motivate actors to share with recipients.
646 Figure 6 plots the effect of anticipated punishment and anger. To illustrate the
647 independent effects of each variable, we perform a median split on differences in anticipated
648 punishment and anger, and divide subjects into four groups accordingly. Figure 8 illustrates that
649 subjects expecting above-median differences in punishment and anger are more likely to share
650 than subjects expecting below-median differences.
651
652
653
654

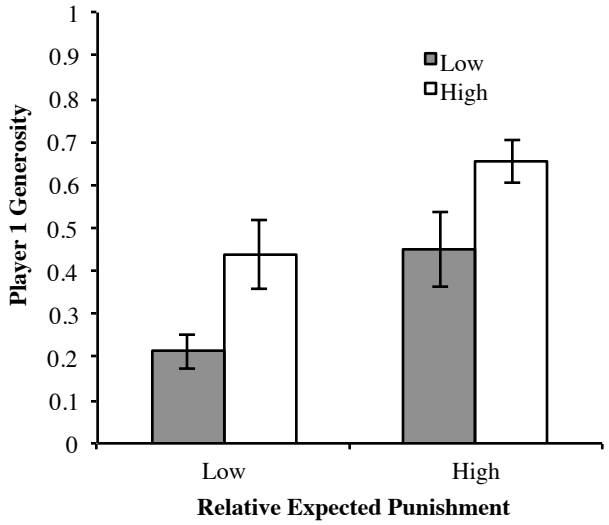


Figure 6. *Actors who expect more third-party punishment and more third-party anger are more likely to choose the fair transfer in Experiment 1. Shown are the fraction of actors choosing the even split, by actor's anticipation of third-party punishment and third-party anger. Relative expected punishment and anger scores are computed by subtracting an individual's expected punishment and anger for fair behavior from their expected punishment and anger for selfish behavior. For ease of visualization, median splits are shown. Error bars indicate robust standard errors of the mean*.

8.2 Discussion of actor and recipient behavior in experiment one

Here, we discuss actor and recipient behavior in experiment one. In experiment one, we find that both actors and recipients expect third parties to punish selfishness behavior more than fair behavior. Further, actors and recipients expect third parties to punish more than they actually do, and to be more sensitivity to actor fairness than they actually are. Finally, actors who anticipate more 3PP and anger in response to selfishness, relative to fairness, are more likely to behave fairly themselves.

These results are consistent with the hypothesis that others expect 3PP, and that expected 3PP promotes cooperative behavior. While much evidence demonstrates that third parties punish selfishness, there is relatively little direct evidence that the possibility of 3PP decreases selfish behavior (an exception is (Charness et al. 2008)). Interestingly, we find that expected anger is

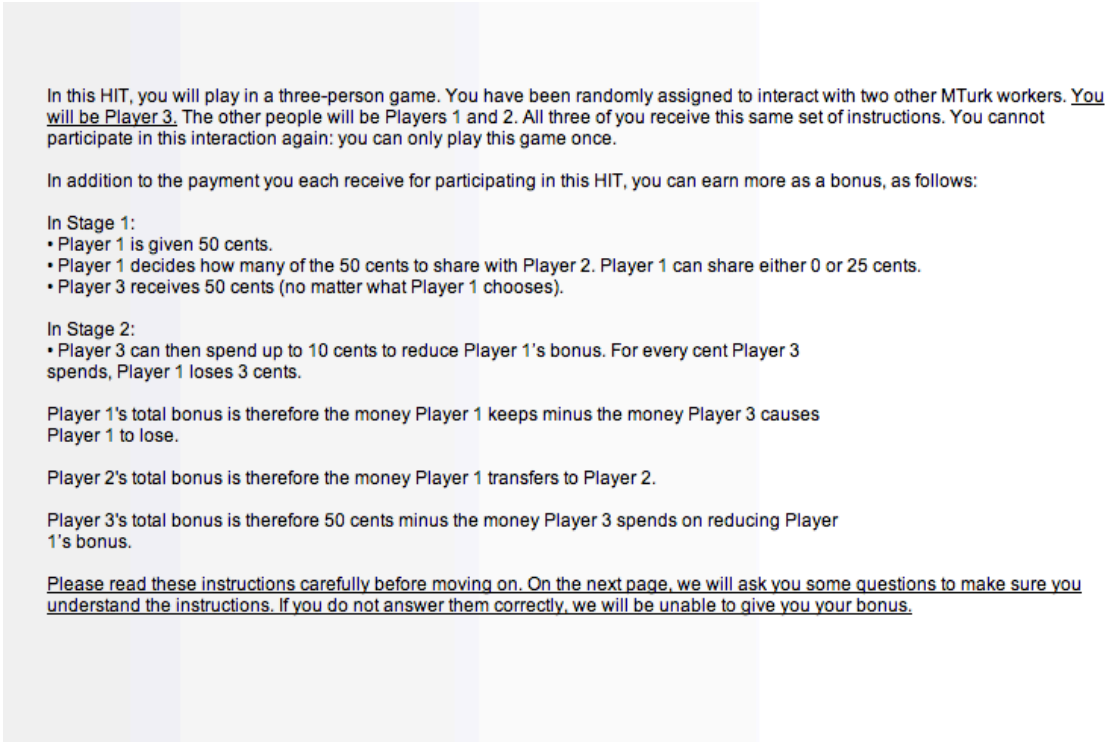
associated with cooperative behavior, over and above the effect of expected punishment. This suggests that third parties may act to promote cooperation not simply through the possibility of material sanctions, but also through the possibility of anger or disapproval.

However, we note that we did not manipulate anticipated 3PP and anger, and the correlational nature of our analysis leaves open the possibility that a third variable caused both increased generosity and anticipated 3PP and anger. For example, subjects with strong other-regarding preferences may be motivated to share in order to increase the welfare of recipients, and also expect third parties to care more about selfishness, experiencing more anger and punishing more harshly.

Interestingly, we did not find that anticipated recipient anger and envy were associated with generosity. One potential explanation is that actors view third-party responses as an unbiased measure of appropriateness, while they expect recipients' responses to reflect self-interest. Alternatively, the possibility for 3PP may crowd out actors' motivation to help recipients (Fehr and Rockenbach 2003; Frey and Jegen 2001). Distinguishing between these possibilities is an interesting direction for future research.

8.3 Experimental instructions

The images below show the instructions and decision screens shown to subjects.



In this HIT, you will play in a three-person game. You have been randomly assigned to interact with two other MTurk workers. You will be Player 3. The other people will be Players 1 and 2. All three of you receive this same set of instructions. You cannot participate in this interaction again: you can only play this game once.

In addition to the payment you each receive for participating in this HIT, you can earn more as a bonus, as follows:

In Stage 1:
• Player 1 is given 50 cents.
• Player 1 decides how many of the 50 cents to share with Player 2. Player 1 can share either 0 or 25 cents.
• Player 3 receives 50 cents (no matter what Player 1 chooses).

In Stage 2:
• Player 3 can then spend up to 10 cents to reduce Player 1's bonus. For every cent Player 3 spends, Player 1 loses 3 cents.

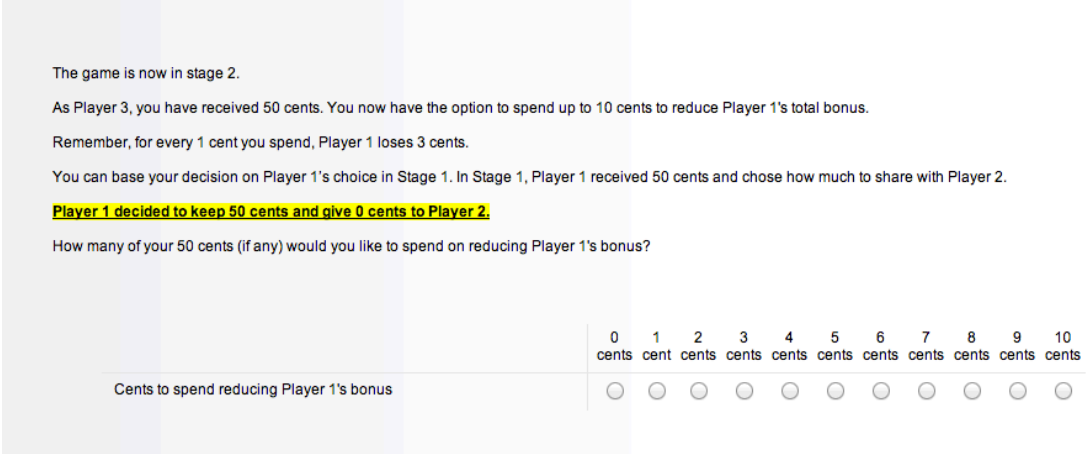
Player 1's total bonus is therefore the money Player 1 keeps minus the money Player 3 causes Player 1 to lose.

Player 2's total bonus is therefore the money Player 1 transfers to Player 2.

Player 3's total bonus is therefore 50 cents minus the money Player 3 spends on reducing Player 1's bonus.

Please read these instructions carefully before moving on. On the next page, we will ask you some questions to make sure you understand the instructions. If you do not answer them correctly, we will be unable to give you your bonus.

Figure 7. *Experiment 1 Instructions (high endowment condition). This image shows the instructions we presented to subjects in the high endowment condition of Experiment 1. Instructions were identical in the low endowment condition, with the exception that they explained that Player 3 receives 25 cents, rather than 50 cents.*



The game is now in stage 2.

As Player 3, you have received 50 cents. You now have the option to spend up to 10 cents to reduce Player 1's total bonus.

Remember, for every 1 cent you spend, Player 1 loses 3 cents.

You can base your decision on Player 1's choice in Stage 1. In Stage 1, Player 1 received 50 cents and chose how much to share with Player 2.

Player 1 decided to keep 50 cents and give 0 cents to Player 2.

How many of your 50 cents (if any) would you like to spend on reducing Player 1's bonus?

	0 cents	1 cent	2 cents	3 cents	4 cents	5 cents	6 cents	7 cents	8 cents	9 cents	10 cents
Cents to spend reducing Player 1's bonus	○	○	○	○	○	○	○	○	○	○	○

Figure 8. *Experiment 1 decision (high endowment, hot condition; selfish offer). This image shows the decisions screen for subjects in the high endowment, hot condition of Experiment 1, who made a decision about a selfish offer. Instructions were identical in the low endowment condition, with the exception that they explained that Player 3 had 25 cents, rather than 50 cents. Instructions were identical for fair actor decisions, with the exception that they said that Player 1 decided to keep 25 cents and give 25 cents to Player 2.*

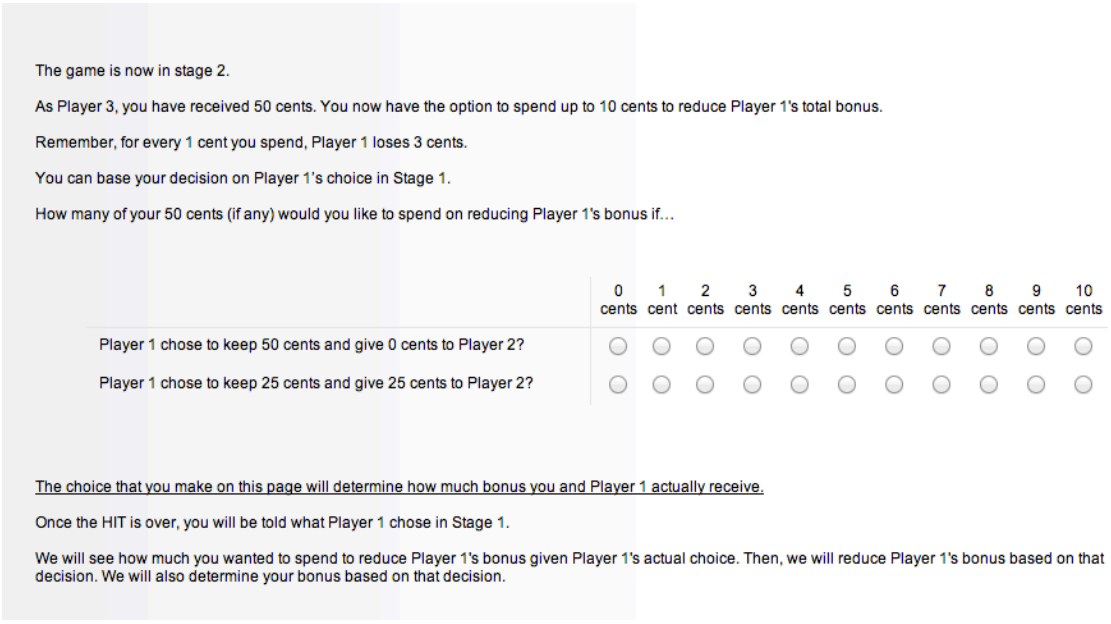


Figure 9. *Experiment 1 decision (high endowment, strategy method condition). This image shows the decisions screen for subjects in the high endowment, strategy method condition of Experiment 1. Instructions were identical in the low endowment condition, with the exception that they explained that Player 3 had 25 cents, rather than 50 cents.*

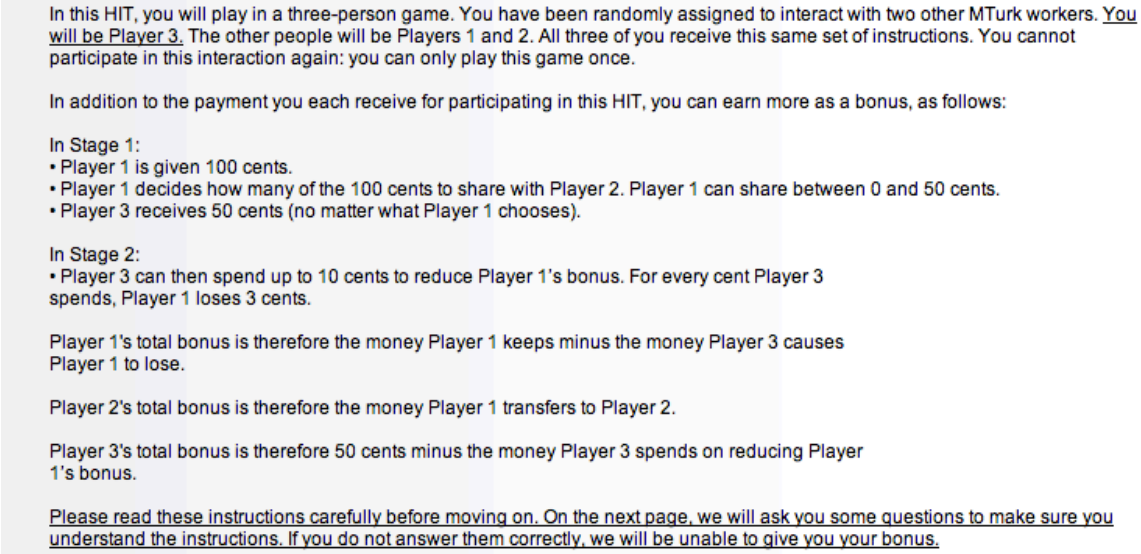


Figure 10. *Experiment 2 Instructions (low endowment condition). This image shows the instructions we presented to subjects in the low endowment condition of Experiment 1. Instructions were identical in the low endowment condition, with the exception that they explained that Player 3 receives 100 cents, rather than 50 cents.*

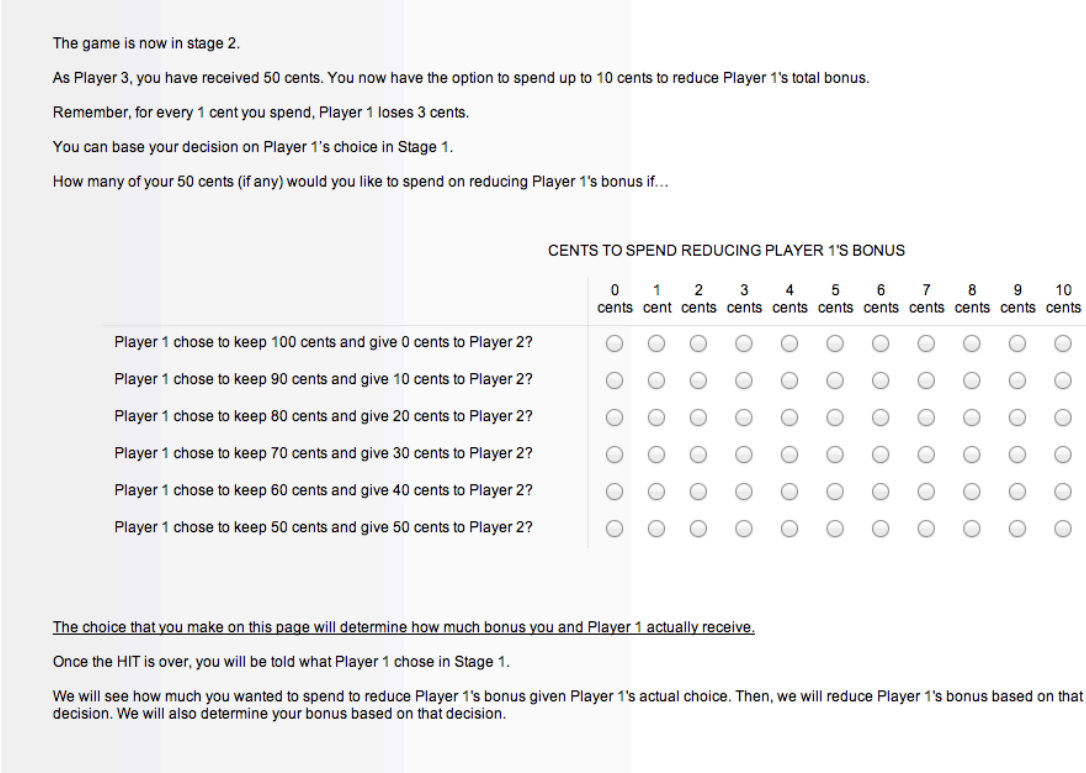


Figure 11. *Experiment 2 decision (low endowment condition). This image shows the decisions screen for subjects in the low endowment condition of Experiment 2. Instructions were identical in the high endowment condition, with the exception that they explained that Player 3 had 100 cents, rather than 50 cents.*

Acknowledgements

We thank Gordon Kraft-Todd for assistance running the experiments, and gratefully acknowledge funding from the John Templeton Foundation.

References

Almenberg, J., Dreber, A., Apicella, C., & Rand, D. (2010). Third party reward and punishment: group size, efficiency and public goods. In *PSYCHOLOGY OF PUNISHMENT, Nova Publishing, Forthcoming*.

Amir, O., Rand, D., & Gal, Y. a. K. (2012). Economic Games on the Internet: The Effect of $1 Stakes. *Plos One, 7*(2), doi:10.1371/journal.pone.0031461.

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters, 122*(2), 308-310.

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

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*(7105), 912-915, doi:10.1038/nature04981.

Bosman, R., & Van Winden, F. (2002). EMOTIONAL HAZARD IN A POWER-TO-TAKE EXPERIMENT. *The Economic Journal, 112*(476), 147-169.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics, 14*(3), 375-398, doi:10.1007/s10683-011-9272-x.

Bruene, M., Scheele, D., Heinisch, C., Tas, C., Wischniewski, J., & Guentuerkuen, O. (2012). Empathy Moderates the Effect of Repetitive Transcranial Magnetic Stimulation of the Right Dorsolateral Prefrontal Cortex on Costly Punishment. *Plos One, 7*(9), doi:10.1371/journal.pone.0044747.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5.

Camerer, C. F., & Hogarth, R. H. (1999). The effects of financial incentives in experiments: A review and capital labor production framework. *Journal of Risk and Uncertainty, 19*, 7-42.

Charness, G., Cobo-Reyes, R., & Jimenez, N. (2008). An investment game with third-party intervention. *Journal of Economic Behavior & Organization, 68*(1), 18-28, doi:10.1016/j.jebo.2008.02.006.

Crockett, M., Apergis-Schoute, A., Herrmann, B., Lieberman, M., Mueller, U., Robbins, T. W., et al. (2013). Serotonin Modulates Striatal Responses to Fairness and Retaliation in Humans. *Journal of Neuroscience, 33*(8), 3505-3513, doi:10.1523/jneurosci.2761-12.2013.

Cubitt, R. P., Drouvelis, M., & Gächter, S. (2011). Framing and free riding: emotional responses and punishment in social dilemma games. *Experimental Economics, 14*(2), 254-272.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica, 73*(6), 2017-2030, doi:10.1111/j.1468-0262.2005.00644.x.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63-87, doi:10.1016/s1090-5138(04)00005-4.

Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature, 415*(6868), 137-140, doi:10.1038/415137a.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*(6928), 137-140.

771 Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation.
772 *Quarterly Journal of Economics*, 817-868.
773 Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy
774 method in public good experiments. *Journal of Economic Psychology, 33*(4), 897-913.
775 Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of economic surveys,*
776 *15*(5), 589-611.
777 Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science,*
778 *317*(5843), 1351-1354, doi:10.1126/science.1144161.
779 Götte, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation
780 and norm enforcement: Evidence using random assignment to real social groups.
781 *American Economic Review, 96*(2), 212-216, doi:10.1257/000282806777211658.
782 Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., et al. (2010).
783 Markets, Religion, Community Size, and the Evolution of Fairness and Punishment.
784 *Science, 327*(5972), 1480-1484, doi:10.1126/science.1182238.
785 Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006).
786 Costly punishment across human societies. *Science, 312*(5781), 1767-1770,
787 doi:10.1126/science.1127333.
788 Horton, J. J., Rand, D., & Zeckhauser, R. J. (2011). The online laboratory: Conducting
789 experiments in a real labor market. *Experimental Economics, 14*(3), 399-425.
790 Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in
791 children's third-party punishment of selfishness. *Proceedings of the National*
792 *Academy of Sciences, 111*(35), 12710-12715.
793 Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment.
794 *Evolution and Human Behavior, 28*(2), 75-84,
795 doi:10.1016/j.evolhumbehav.2006.06.001.
796 Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of economic*
797 *dynamics, 1*(3), 593-622.
798 Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., et al.
799 (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal*
800 *Society B-Biological Sciences, 275*(1634), 587-590, doi:10.1098/rspb.2007.1517.
801 Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk.
802 *Behavior research methods, 44*(1), 1-23.
803 McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young
804 children. *Cognition, 134*, 1-10.
805 Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party
806 punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and*
807 *Decision Making, 4*(7), 543-553.
808 Nikiforakis, N., & Mitchell, H. (2013). Mixing the carrots with the sticks: third party
809 punishment and reward. *Experimental Economics*, 1-23.
810 Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon
811 mechanical turk. *Judgment and Decision Making, 5*(5), 411-419.
812 Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish
813 altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences,*
814 *280*(1758).
815 Rand, D. (2012). The promise of Mechanical Turk: How online labor markets can help
816 theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172-179.

817 Rand, D., Arbesman, S., & Christakis, N. A. (2011). Dynamic social networks promote
818 cooperation in experiments with humans. *Proceedings of the National Academy of*
819 *Sciences, 108*(48), 19193-19198.
820 Rand, D., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed.
821 *Nature, 489*(7416), 427-430.
822 Selten, R. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens
823 im Rahmen eines Oligopolexperimentes. In*, 1965*: Seminar für Mathemat.
824 Wirtschaftsforschung u. Ökonometrie
825 Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter
826 enemies:“Altruistic” punishment of in-group members. *Evolution and Human*
827 *Behavior, 25*(6), 379-393.
828 Suri, S., & Watts, D. J. (2011). Cooperation and Contagion in Web-Based, Networked Public
829 Goods Experiments. *PLoS ONE, 6*(3), e16836.
830 Watson, D., & Clark, L. A. (1991). Self-Ratings Versus Peer-Ratings of Specific Emotional
831 Traits - Evidence of Convergent and Discriminant Validity. *Journal of Personality and*
832 *Social Psychology, 60*(6), 927-940, doi:Doi 10.1037//0022-3514.60.6.927.
833 Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief
834 Measures of Positive and Negative Affect - the Panas Scales. *Journal of Personality*
835 *and Social Psychology, 54*(6), 1063-1070, doi:Doi 10.1037/0022-3514.54.6.1063.
836