October 2006

Audience Effects on Moralistic Punishment

Robert Kurzban
University of Pennsylvania

Peter DeScioli
University of Pennsylvania

Erin O'Brien
University of Pennsylvania

ELSEVIER

Evolution and Human Behavior xxx (2006) xxx–xxx

Evolution and Human Behavior

Audience effects on moralistic punishment☆

Robert Kurzban*, Peter DeScioli, Erin O'Brien

Q1 *Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA*

Abstract

Punishment has been proposed as being central to two distinctively human phenomena: cooperation in groups and morality. Here we investigate moralistic punishment, a behavior designed to inflict costs on another individual in response to a perceived moral violation. There is currently no consensus on which evolutionary model best accounts for this phenomenon in humans. Models that turn on individuals' cultivating reputations as moralistic punishers clearly predict that psychological systems should be designed to increase punishment in response to information that one's decisions to punish will be known by others. We report two experiments in which we induce participants to commit moral violations and then present third parties with the opportunity to pay to punish wrongdoers. Varying conditions of anonymity, we find that the presence of an audience—even if only the experimenter—causes an increase in moralistic punishment.

Keywords: Punishment; Altruism; Reciprocity; Cooperation; Reputation

1. The evolution of moralistic punishment

People punish wrongdoers, intervening even when they themselves have not been harmed. Third-party punishment (TPP) has been observed in the field (Sober & Wilson, 1998) and in

☆ Portions of this research were made possible by a Nassau Fund grant from the Center for Undergraduate Research and Fellowships at the University of Pennsylvania to Erin O'Brien and a MacArthur Foundation grant to Robert Kurzban.

* Corresponding author. Tel.: +1 215 898 4977; fax: +1 215 898 7301.
E-mail address: kurzban@psych.upenn.edu (R. Kurzban).

1090-5138/06/$ – see front matter © 2006 Published by Elsevier Inc.
doi:10.1016/j.evolhumbehav.2006.06.001

the laboratory (e.g., Fehr & Fischbacher, 2004), and is a crucial feature of human social life, forming the cornerstone of morality (e.g., Wilson, 1993; Wright, 1995). Humans everywhere seek and assess evidence of infractions, identify acts as morally right or wrong, and desire that wrongdoers be punished (Brown, 1991). We regard moralistic punishment as a behavior caused by systems designed to inflict costs in response to wrongdoing.

Among nonhuman animals, punishment is typically confined to interactions in which individuals have a direct interest. There are, however, several putative exceptions. Chimpanzees have been observed to intervene on behalf of unrelated others (de Waal, 1996), macaques punish conspecifics that fail to announce the finding of food (Hauser & Marler, 1993), and several ant species attack and kill rogue workers attempting to lay their own eggs (e.g., Gobin, Billen, & Peeters, 1999).

Moralistic punishment in humans is an evolutionary mystery because it is performed by third parties. This raises the key question: Why do people care about interactions among unrelated others? Given that punishment is costly and can potentially draw retaliation, TPP appears to be a tendency that would be selected against, raising the issue of how adaptations that give rise to moralistic punishment evolved.

2. Models of the evolution of moralistic punishment

Punishment has been linked with the evolution of cooperation in groups (Boyd & Richerson, 1992)—a connection that has strengthened in recent years (Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Gächter, 2002). Briefly, cooperation in groups of unrelated individuals is difficult to explain because individuals stand to gain by enjoying the benefits of group efforts without contributing (i.e., "free riding"). Punishment is a frequently proposed solution because, if sufficient costs are inflicted on free riders, then cooperators are at a selective advantage (Fehr & Gächter, 2002). However, because punishing noncooperators itself entails a cost, nonpunishers in a group possess a relative advantage, making the evolution of punishment itself problematic (see, e.g., Boyd et al., 2003).

One potential resolution is that punishment might have evolved as a result of group benefits, despite costs to punishing individuals. By curtailing free riding, groups with punishers might outcompete groups without punishers. One important example is the model of strong reciprocity (Fehr, Fischbacher, & Gächter, 2002; Gintis, 2000, 2005). According to Gintis (2000), "[a] strong reciprocator is predisposed to cooperate with others and punish noncooperators, even when this behavior cannot be justified in terms of self-interest, extended kinship, or reciprocal altruism" (p. 169).

Other models imply that moralistic punishment is designed to benefit the individual by virtue of its effects on others' perceptions. Johnstone and Bshary (2004), for example, have shown that indirect reciprocity can favor costly punishment when these acts discourage future aggression by observers. More generally, cognitive mechanisms underlying moralistic punishment might have evolved because of their signaling benefits. It is well known that costly and seemingly inefficient morphological or behavioral traits can be favored by natural selection as honest signals of quality (Zahavi, 1975). Costly signals can yield a fitness advantage when they reliably correlate with underlying traits that are difficult to observe, such as one's quality as a mate, ally, or exchange partner (for an extended discussion, see Miller, 2000). A reliable correlation between signal and quality is obtained when higher quality individuals face lower costs or higher benefits associated with the signal. Under these conditions, adaptations for both signaling and receiving the signal can be favored by selection.

Indeed, Gintis, Smith, and Bowles (2001) found that punishment can yield signaling benefits when high-quality individuals have reduced costs or increased benefits associated with punishment. If this explanation is correct, moralistic punishment constitutes an advertisement of individual quality, selected by virtue of the reputational advantages it confers.[1] Similarly, Fessler and Haley (2003) have suggested that moralistic punishment is designed to signal that one is a good candidate for cooperative interaction because it demonstrates knowledge of, and support for, local behavioral norms (see also Barclay, in press).

Models driven by reputation effects, such as costly signaling, predict adaptations designed to influence others' representations. That is, these models imply that selection pressures favored cognitive mechanisms whose operation is mediated by the presence of an audience. To the extent that any costly behavior functions to alter others' perceptions, underlying cognitive systems should be sensitive to the presence of others (Burnham & Hare, in press; Haley & Fessler, 2005). Therefore, based on these models, we should expect to find evidence that moralistic punishment is sensitive to social presence (e.g., Fessler & Haley, 2003; for a nice treatment of recent work and relevant theory, see also Carpenter, in press; Carpenter & Matthews, 2004). The experiments described here investigate the proximate mechanisms that underpin moralistic punishment, which might in turn help to illuminate ultimate explanations.

3. Previous work

Two lines of previous research are relevant to the current question: (a) studies of TPP, and (b) studies investigating how cues to social presence affect decisions in strategic interactions. Experimental economists have been interested in costly punishment in large measure because it constitutes a violation of self-interest when punishment cannot be a deterrent in future interactions, as in the one-shot Ultimatum Game (for a recent review, see Camerer, 2003). Our interest extends into the more specific domain of moralistic punishment. Our focus on audience effects makes relevant the effects of the presence of other people, or simply cues to their presence.

3.1. Do people engage in TPP?

In an early experiment on TPP (Kahneman, Knetsch, & Thaler, 1986), participants endured a cost 74% of the time to reduce the payment of participants who chose an uneven split

[1] We leave aside the issue of whether and why people tend to want to punish actions that are detrimental to their groups (Boyd & Richerson, 1992). This issue is important but is beyond the scope of this paper.

(i.e., were "unfair") in a Dictator Game. However, punishing unfair players and rewarding fair players were confounded in this study. Subsequently, Turillo, Folger, Lavelle, Umphress, and Gee (2002) removed this confound and found that only 15% of their participants punished unfair players—a proportion not significantly different from the proportion of individuals who punished fair players (see also Ottone, 2004).

Most closely related to the studies reported here, Fehr and Fischbacher (2004) examined TPP in the context of one-shot Dictator Game and Prisoner's Dilemma Game. In the TPP dictator experiment, Player A transferred 0–100 points (in increments of 10) to Player B. An uninvolved Player C indicated, for every level of Player A's transfer, how much of their 50-point endowment they would spend to reduce Player A's payoff, each point resulting in a three-point reduction. More than 60% (14 of 22) of participants were willing to pay to punish. When dictators transferred nothing, third parties spent an average of 14 points (28% of their endowment) on punishment, although dictators nonetheless profited from selfishness. In the analogous Prisoner's Dilemma Game, defectors were punished most severely when the defector's counterpart cooperated. In this case, 46% (11 of 24) of third parties punished defectors, and the overall average expenditure on punishment was 3.35 points (8.4% of endowment).[2]

TPP has also been investigated in the context of public goods games (Ledyard, 1995) in which people in one group are able to inflict costs on members of another group. Carpenter and Matthews (2005) found that only 10% of participants punished individuals in a group different from their own, and the overall amount spent to punish individuals in a different group was about US$0.10—a small amount given the average earnings of US$16 (net of show-up payment) per participant.

In sum, the TPP documented in previous studies ranged in magnitude from negligible to modest. Questions remain, however, about the role of anonymity.

3.2. Cues to social presence in economic games

The effect of the presence of others has a long and distinguished history in social psychology, dating back at least as far as early work on "social facilitation" (Zajonc, 1965; see also Triplett, 1898). Effects of observation are influenced by task difficulty (Markus, 1978), the extent to which one's performance is being evaluated (Cottrell, Wack, Sekerak, & Rittle, 1996), and details about the observer (Butler & Baumeister, 1998). The presence of others has long been known to have effects on decisions to engage in more prosocial (Latane, 1970)—and less antisocial (Diener, Fraser, Beaman, & Kelem, 1976)—behavior, consistent with the view that people are concerned about others' perceptions of them, especially in the domain of morality (Jones & Pittman, 1982).

[2] From the manuscript and from instructions to participants, it is not possible to know what participants believed regarding the experimenter's knowledge of their decisions.

Of particular relevance, Hoffman, McCabe, Shachat, and Smith (1994) found that, in a Dictator Game, when participants are assured that the experimenter will not know how much money they choose to transfer, the majority of participants give US$0, less than what is typically found in such games (e.g., Forsythe, Horowitz, Savin, & Sefton, 1994). Furthermore, as predicted by modular approaches, cues that one is being observed increase prosocial behavior, even in the absence of actual observation (Kurzban, 1998; for a recent extended discussion of modularity, see Barrett & Kurzban, in press). Kurzban (2001), for example, showed that, in a public goods game, having people exchange mutual oblique eye gazes (but no information about others' contributions) increased contributions to the public good in (all-male) groups compared to a control condition with no eye gaze. Burnham and Hare (in press) and Haley and Fessler (2005) have shown similar effects of cues on social presence in a Dictator Game and in a public goods game, respectively.

4. Current studies: hypotheses and predictions

The experiments reported below investigate the role of social presence on decisions to punish moral violations—in this case, expectations of trust and reciprocity. In the first stage, we use the "Trust Game" (Berg, Dickhaut, & McCabe, 1995) and the Prisoner's Dilemma Game to elicit norm-violating behavior. We then allow participants in the second stage to pay to inflict costs on individuals who have acted "untrustworthily" (Experiment 1) or on individuals who have failed to reciprocate a cooperative move (Experiment 2).

In both experiments, we manipulate participants' beliefs regarding who will know their decisions to punish. In the "Anonymous" condition, participants are led to believe (truthfully) that no one, including the experimenter, will know how much any particular participant chose to punish. We reason that punishment under these circumstances cannot be attributed to (conscious) concerns for garnering a reputation for punishing defectors. In our Treatment conditions, participants are led to believe (again, truthfully) that others will know how much they have chosen to punish. On the basis of previous results and of the broad literature on the importance of self-presentational motives (e.g., Kurzban & Aktipis, in press), we predict that TPP will be minimal under conditions of anonymity but will be substantially greater when participants are observed.

5. Experiment 1: TPP in a Trust Game

5.1. Method

5.1.1. Participants

Fifty-eight undergraduates were recruited at the University of Pennsylvania through the "Experiments @ Penn" web-based recruitment system. Participants were told that they would earn a participation payment for showing up and could earn additional money depending on decisions made during the experiment. To make participants feel less identifiable, no demographic information was collected in this experiment.

5.1.2. Procedure

The experiment was conducted in two stages: a Trust Game (Berg et al., 1995) played by one group of participants in a morning session, and a subsequent "punishment" round played by a different set of participants in a set of afternoon sessions. Five experimental sessions were held in the Penn Laboratory for Evolutionary Experimental Psychology (PLEEP) at the University of Pennsylvania. This laboratory consists of 16 stations divided by partitions. All decisions were made by pencil and paper, and all participants participated in one stage and one session only. (Complete instructions and experimental materials are available on request.)

The first stage was designed to elicit a violation that would be perceived as warranting punishment.[3] Fourteen participants played a Trust Game. Participants were randomly assigned as decision maker 1 (DM1) or decision maker 2 (DM2). An index card with a series of identity-masking codes was placed at the kiosks at which the participants were seated. Decision makers would be paid for one interaction, with each one identified by one code on this index card. When participants returned to the laboratory at the end of the day, a code on this index card would be matched to a code on an envelope containing the participant's payment. Participants were paid their US$5 show-up payment at the end of the session and paid additional earnings when they returned at 1700 h on that day.

DM1 received five game pieces with five extensive-form Trust Games. DM1s could move right, ending that particular game and splitting US$20 with DM2, or could move down, thereby "trusting" DM2. If DM1 moved down, DM2 decided between the outcome (US$20, US$20), the "trustworthy" choice, and the "untrustworthy" choice. The untrustworthy payoffs varied across the five games and were (US$12, US$28), (US$9, US$31), (US$6, US$34), (US$3, US$37), and (US$1, US$39). After DM1s had made decisions in all five games, game pieces were collected, shuffled, and distributed to DM2s. DM2s wrote their subject codes on all game pieces and indicated their choices when applicable.

All decisions were made anonymously; choices were identified by subject codes, and game pieces were concealed in envelopes to ensure anonymity. The first-stage session lasted 45 min. Written instructions directed participants to retain their index card with subject codes and to return later in the afternoon to receive their payment. Instructions to DM2s indicated that decisions made by participants in later sessions could affect their payment, although no additional details regarding how their payment could be affected were given (see footnote 3). Participants were paid based on one of the five games they played, possibly reduced by a punishment from subsequent sessions (see below). Stage 1 participants earned an average of US$17.50, including the US$5 show-up payment. All participants returned to claim their earnings.

From this first stage, one game piece on which DM1 had chosen to move down and DM2 had chosen to move right was selected, reaching the US$1/US$39 outcome. This game piece was photocopied, and one copy was used for all subsequent punishment decisions in the

[3] Fabricating a norm-violating play would have simplified matters. However, we followed the norms in behavioral economics and eschewed the use of deception (Hertwig & Ortmann, 2001). The PLEEP laboratory and the Experiments @ Penn web-based recruiting system have a policy against deception. The method used here pushes the envelope of nondeception. However, nothing false was told to participants.

second stage of the experiment. For all other DM1s and DM2s, one interaction was randomly selected and payoffs were computed.

In the second stage, a new set of people participated in a punishment phase. Participants were presented with instructions and with tasks that participants in the morning session had completed. In addition, they were given a photocopy of the game piece from the first stage of the experiment in which DM1 moved down and DM2 chose the maximally selfish outcome (US$1, US$39). Participants were given US$7 and instructed that they could spend none, some, or all of this money, in US$1 increments, to be deducted from DM2's payment. The remaining money was theirs to keep. Each dollar spent reduced DM2's payoff by US$3, allowing reductions of US$0–21. These punishments were averaged to compute the amount deducted from this particular DM2. The instructions did not use the term *punish* or *sanction*, but used instead the more neutral term *deduction*.

5.1.3. Treatments

There were two conditions: Anonymous[4] (n=24) and Experimenter (n=19). In the Anonymous condition, participants divided their US$7 into two envelopes (one for deduction and one for themselves), which were color-coded for distinguishability. After making their decision, the participants, one at a time, took both sealed envelopes with them as they left the room. Outside the room was an opaque bin with a narrow slit into which an envelope could be dropped. Participants were instructed to drop their sealed deduction envelopes into this bin as they left the experiment, taking the remaining envelope with them.[5] Participants were told, truthfully, that it would be impossible for anyone to know how much they spent on punishment. Although the policy at PLEEP is that participants will not be deceived, we cannot verify independently that this policy itself is known and believed to be true by our participants.

In the Experimenter condition, participants were informed that their decision would be known to the experimenter. In particular, they would meet an experimenter outside the laboratory where they would count the amount spent to reduce the payoff to DM2. Two sessions in each condition were conducted. The sessions lasted 30 min, and participants earned an average of US$8.48, including the US$3 participation payment.

5.2. Results

Overall, DM1s chose not to trust DM2 in 60% (21 of 35) of cases. Conditional on DM1 moving down, DM2s chose the uneven outcome in 64% (9 of 14) of cases. When the payoffs (DM1, DM2) were (US$1, US$39), (US$3, US$37), and (US$6, US$34), only one of seven DM1s moved down ("trust"); in each case, DM2 chose the uneven split. When the payoffs were (US$9, US$31), four of seven DM1s "trusted," and only one DM2 proved to be

[4] We use the term *anonymous* rather than "double-blind" because experimenters were not blind to treatment conditions.

[5] Preserving anonymity while retaining the ability to gather individual (as opposed to aggregate) data is a nontrivial methodological challenge [for a discussion see Bolton & Zwick, 1995; for a sense of the intricacies of such procedures, which they describe as "quite involved" (p. 273), see Bolton, Katok, & Zwick, 1998, especially their Fig. 2].

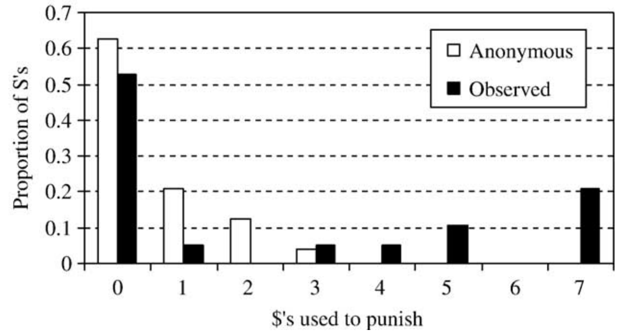

Fig. 1. Distribution of punishment decisions in Experiment 1. (The outlier has been omitted; see text.)

trustworthy. When the payoffs were (US$12, US$28), six of seven DM1s trusted, and three DM2 proved to be trustworthy. These results are peripheral to our method. The single trusting move by one DM1 when the payoffs were (US$1, US$39), with the subsequent untrustworthy move by DM2, generated the stimulus object needed for the subsequent punishment round.

Central to our hypotheses is the behavior of participants in the session in which DM2s could be punished.[6] In the punishment round, 38% (9 of 24) of third-party participants paid to punish in the Anonymous condition, while 47% (9 of 19) punished in the Experimenter condition. Because of skewed distribution and directional prediction, we conducted a one-tailed Wilcoxon rank-sum test (z=1.52, p=.06), obtaining a result just shy of standard levels of significance. We therefore conducted an additional analysis that retains some of the information lost in the Wilcoxon test. We treated punishment as a binary variable, categorizing each punishment decision (Fig. 1) as either less than half of the endowment (US$0–3) or greater than half of the endowment (US$4–7). Using this test, the difference between the two conditions is statistically significant (p=.002, Fisher's Exact Test). This result is still significant after a Bonferroni correction for all six possible divisions between zero and seven (adjusted α=.05/6=.008).

5.3. Discussion

People punished more when their decision would be known to the experimenter than under conditions of anonymity. This result is consistent with reputation-based accounts of moralistic punishment.

The Trust Game, however, might not be the best means of eliciting a "norm violation." Indeed, researchers in the behavioral economics literature differ on the interpretation of

[6] One individual in the Anonymous condition asked the experimenter questions that indicated thorough confusion and revealed that they had chosen to punish the maximum amount. Because they did not understand the task and informed the experimenter of their decision (thus reassigning themselves from the Anonymous treatment to the Experimenter treatment), we proceeded with our analysis, omitting this observation.

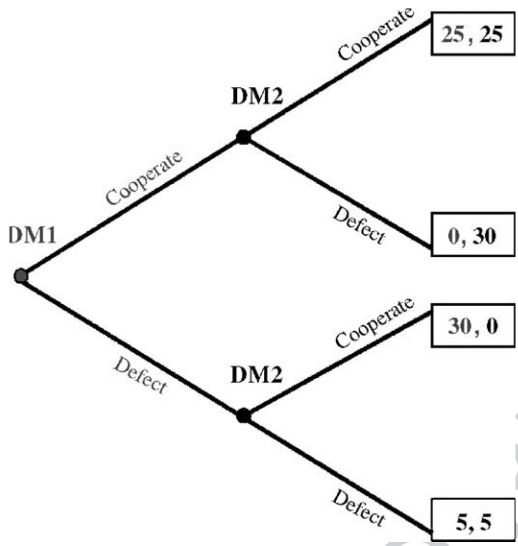

Fig. 2. The sequential Prisoner's Dilemma Game (in extensive form) used in Experiment 2. For payoff information, see Section 6.1.2.

decisions in the Trust Game (see, e.g., Cox, 2004; Cox & Deck, 2005), and it is not clear that our participants uniformly construed DM2's move to the right as untrustworthy. This raises questions about both the use of the Anonymous condition as an index of a taste for punishment and the use of the Treatment condition as an index of a desire for a positive reputation. In Experiment 2, we used a sequential Prisoner's Dilemma Game to obtain less ambiguous norm violations.

6. Experiment 2: TPP in a Prisoner's Dilemma Game

Experiment 2 used a sequential Prisoner's Dilemma Game in extensive form. The sequential game was used because defection following cooperation is very naturally interpreted as a violation of reciprocity (cf., McCabe, Smith, & LePore, 2000; Schotter, Weiss, & Zapater, 1996). We also labeled the edges in the extensive-form game with the words "Cooperate" and "Defect" to maximize the chance that all participants construed these decisions in the same way (see Fig. 2). Finally, we added a condition in which not only the experimenter but also other participants would know the extent to which participants chose to punish defectors. This additional treatment helps to determine whether the number of observers influences decisions to punish as a third party.

6.1. Method

6.1.1. Participants

One hundred three (72 female, 31 male) undergraduates were recruited at the University of Pennsylvania through the Experiments @ Penn electronic recruitment system. All participants were at least 18 years of age, with a mean (S.D.) age of 21 (3) years, and all were fluent English speakers. In a departure from the procedure in Experiment 1, because we judged that adding demographic items would not undermine participants' sense of anonymity, we asked participants to indicate their age and sex on a short questionnaire after they had made their decisions. Participants were told that they would earn a US$5 participation payment for showing up and could earn additional money depending on decisions made during the experiment.

6.1.2. Procedure

Experiment 2 largely replicated Experiment 1, with minor modifications. The sequential Prisoner's Dilemma Game substituted for the extensive-form Trust Game. Seven experimental sessions were held. In Stage 1, 16 participants played the one-shot sequential Prisoner's Dilemma Game. Participants were randomly assigned as DM1 or DM2. The five games had different payoffs. In all games, Cooperate–Defect yielded (US$0, US$30), and Defect–Cooperate yielded (US$30, US$0). Cooperate–Cooperate and Defect–Defect payoffs varied, with the payoffs for mutual cooperation/defection, respectively, as follows: (US$25, US$5), (US$20, US$5), (US$18, US$5), (US$16, US$5), and (US$18, US$7). DM1 received five game pieces depicting the games (see Fig. 2). Subject codes and decisions were indicated on each game piece. After DM1s had made decisions to Cooperate or Defect in all five games, game pieces were collected, shuffled, and distributed to DM2s. DM2s then chose whether to Cooperate or to Defect, determining the final outcome of the game.

The procedures for maintaining anonymity and for paying participants were identical to those used in Experiment 1. Stage 1 participants earned an average of US$13.80, including the US$5 participation payment. In the second stage, a different set of participants could pay to punish selfish DM2s from the first stage. Participants were given a photocopy of a game piece and instructions from the first stage of the experiment in which DM1 chose Cooperate and DM2 chose Defect. The game piece selected for punishment was the one in which mutual cooperation would have yielded symmetrical payoffs of US$25 each. Instead, the Cooperate/Defect outcome yielded payoffs of (US$0, US$30).

Participants were given US$5 as their show-up payment (an endowment of US$10 in US$1 bills) and were able to use US$0–10 to be deducted from DM2's payment, while the remaining money was theirs to keep. Each dollar spent reduced DM2's payoff by US$3, allowing reductions of US$0–30, which could potentially reduce DM2's payoff to US$0. As in Experiment 1, only the term *deduction* was used in the instructions, and punishments were averaged to compute the amount deducted from this particular DM2.

6.1.3. Treatments

There were a total of six Stage 2 sessions, two in each of three experimental conditions: Anonymous, Experimenter, and Participants (n=31, 26, and 30, respectively). The Anonymous and Experimenter conditions were identical to the treatments in Experiment 1. In the Participants condition, participants were informed that, after everyone had made his/her decision and had sealed his/her envelope (to prevent changes), each participant would be asked to stand and announce the outcome of the game piece (i.e., "Cooperate–Defect") and

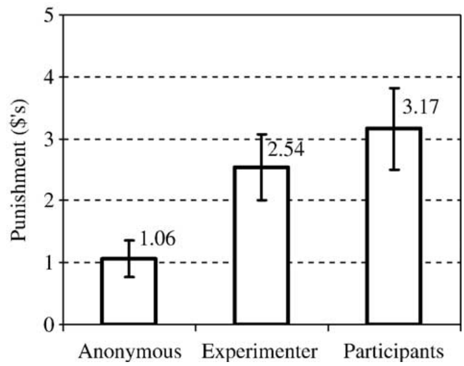

Fig. 3. Punishment decisions in Experiment 2. Error bars are 1 S.E. The full scale is not shown. The total punishment possible is US$10.

the amount that they spent on punishment. Participants were told that their decision would be known to all participants in the session and to the two experimenters. Because the size of the audience might be important, we note that the number of participants was n=14 and n=16 in Sessions 1 and 2, respectively. The sessions lasted 30 min, and participants earned an average of US$12.77, including the US$5 show-up payment.

After making their decisions, participants were asked to fill out a short survey that asked about the reasoning behind their allocation decision.

6.2. Results

In the sequential Prisoner's Dilemma Game, (Cooperate, Cooperate), (Cooperate, Defect), (Defect, Cooperate), and (Defect, Defect) occurred 6, 8, 10, and 16 times, respectively. The relatively high frequency of (Defect, Cooperate) is extremely unusual, a result for which we have no good explanation. It is, however, irrelevant to the present study, as the Prisoner's Dilemma Game was used only to generate a Cooperate–Defect sequence of moves.

The proportion of participants who engaged in costly punishment in the Anonymous, Experimenter, and Participants conditions was 42% (13 of 31), 65% (17 of 26), and 67% (20 of 30), respectively. The mean (S.D.) expenditure on punishment was US$1.06 (1.65), Q3 US$2.54 (2.70), and US$3.17 (3.60), respectively (Fig. 3).

Again because of the distribution of the data, we conducted a nonparametric Kruskal–Wallis rank-sum test, finding that money spent on punishment differed across conditions [$\chi^2(2, N=87)=7.56$, $p=.02$]. We further conducted pairwise Wilcoxon rank-sum tests, which showed that more money was spent on punishment in the Experimenter condition than in the Anonymous condition ($z=2.25$, $p=.02$), and that more money was spent on punishment in the Participants condition than in the Anonymous condition ($z=2.47$, $p=.01$). Punishment did not differ significantly between the Experimenter and Participants conditions ($z=.33$, $p=.74$). Selfish individuals gained US$5 by defecting, while they incurred average punishments of US$3.18, US$7.62, and US$9.51 in the Anonymous, Experimenter, and Participants conditions, respectively.

Table 1
Experiment 2: reliability and mean ratings (S.D.) of free responses

		Condition		
Scale	Cronbach's α	Anonymous	Experimenter	Participants
Angry	.89	1.86 (0.70)[a]	2.45 (1.45)[b]	2.39 (1.26)[b]
Disgusted	.87	1.88 (0.75)[a]	2.28 (1.34)[a,b]	2.47 (1.38)[b]
Contemptuous	.89	1.98 (0.77)[a]	2.53 (1.38)[a,b]	2.64 (1.37)[b]
Selfish	.88	4.83 (1.20)[a]	3.97 (1.70)[a,b]	3.46 (1.31)[b]

Ratings are on a scale from 1 to 7 (see text). Within each row, entries that do not share a superscript differ at $p<.05$. For "selfish," the difference between Anonymous and Participants is significant at $p<.0001$. Because we predicted greater emotion in the Experimenter and Participants conditions, and greater selfishness in the Anonymous condition, all tests are one-tailed.

Q4 Three independent raters scored participants' comments explaining their decision on a scale from 1 to 7. Raters were asked to indicate "how [X] the person making comments seems to be," where X=angry, disgusted, contemptuous, guilty, ashamed, and selfish. Because Cronbach's α (a measure of interrater agreement)values for guilty and ashamed were only .69 and .67, respectively, we omitted these results (see Table 1).

6.3. Discussion

Under conditions of anonymity, participants punished someone who defected after a cooperative move in a sequential Prisoner's Dilemma Game, but this punishment was small—roughly US$1 or 10% of the possible amount they could punish. Knowledge that the experimenter, or the experimenter and other participants were going to know how much an individual punished increased this amount—more than tripling it in the latter case.

Quite unexpectedly, in the Participants condition, at least one subject attempted to deceive others by announcing a false outcome. Because we did not anticipate deception, we did not record this information and could not determine the relationship between dissembling and punishment decisions. We suspected, but could not confirm, that those who punished the least were most likely to attempt deception. The fact that we observed dissembling testifies to the importance of computations regarding reputation.

7. Conclusion

Perhaps the best summary of our results comes from one participant in Experiment 2 (Anonymous condition): "Since it's anonymous, [there is] no reason not to take as much money as I could. But [I] figured I should start deducting at least a little from DM2." This is consistent with our broad results from both experiments. Under Anonymous conditions, people did punish, but relatively little. Some normative motive, indicated by the modal "should," might be at work.

In contrast, punishment increased even when only one person knew the decisions made by the participant. In the presence of roughly a dozen participants, punishment expenditure tripled. Of course, participants probably did not expect to encounter audience members again, suggesting that the effect is driven by social presence per se rather than by conscious computations associated with interacting with that particular individual again, consistent with findings described in Section 3.2. No participants indicated in their free responses that they were punishing because they were being observed. This implies either additional self-presentational concerns (not wanting to appear to be punishing only because they are being watched) or a genuine lack of knowledge of their own motives (Nisbett & Wilson, 1977), consistent with the theory surrounding modularity (Hirstein, 2005; Kurzban & Aktipis, in press).

Self-report data from the second experiment suggest the action of two separate mechanisms. Participants in the nonanonymous conditions reported greater anger and less selfishness (see also Elster, 1998; Ketelaar & Au, 2003). This suggests that observations might activate emotional systems (e.g., anger) and attenuate systems for computing one's own economic interest. Because these effects were relatively small and derived from self-report, caution should be exercised in interpreting them.

7.1. Situating the results

The implications of our results for evaluating relevant theory can be seen most clearly in the context of work on the effect of anonymity in a slightly modified version of the Ultimatum Game. Bolton and Zwick (1995) used a set of extensive-form games in which the first decision maker could choose to allow the second decision maker to select between one of two options: (a) US$2 for each or US$0 for each, or (2) an unequal split (benefiting DM1) of US$4 (e.g., US$3.40/US$0.60) and US$0 for each. In the latter case, the choice of US$0 for each is interpretable as a punishment for DM1 choosing to forgo the possibility of evenly splitting the US$4 endowment. In a condition analogous to our anonymity treatment, in which DM2's decisions were unknowable by the experimenter, anonymous punishment of uneven splitters was very similar to the control condition [see especially Figs. 5 and 6 (pp. 110 and 111, respectively) of Bolton & Zwick, 1995], Bolton and Zwick conclude that the effect of being observed by an experimenter is "relatively weak" (p. 113) compared to the "propensity to punish those who treat them 'unfairly,' independent of any influence exerted by the experimenter" (p. 96).

These results, combined with those from the present study, suggest that anonymity has a weaker effect in the context of second-party punishment than in the context of TPP. This speaks of the question on the nature of psychological design and the ultimate explanation for design features. Putting simply, these results raise the possibility that punishing someone who has treated you unfairly is a taste that can override the taste for individual gain or wealth, and is not substantially mediated by cues that one is being observed. In contrast, the taste for TPP is weak compared to the taste for individual gain, and is mediated by cues that one is being observed.

Bridging from these results to ultimate explanations must necessarily be tentative. However, this contrast hints that adaptations for second-party punishment might have been driven by selection pressures associated with repeat interactions with particular individuals

(Trivers, 1971). In contrast, adaptations for TPP might have been driven, at least in part, by selection pressures associated with reputation, as suggested by sensitivity to observation. The small amount of TPP under conditions of anonymity is subject to a wide variety of interpretations, including "mismatch" explanations (Hagen & Hammerstein, in press) and the models described above (e.g., Gintis, 2000). Future work will need to clarify the design features associated with both types of punishment. The current data raise the possibility of different histories of selection for the computational systems that underpin these two types of punishment, and that they might be, to some extent, functionally distinct.

7.2. Are demand characteristics an alternative explanation?

Demand characteristics refer to features of an experiment that allow participants to infer what is expected of them and thereby cause them to act in that way, limiting the inferences that can be drawn from the experiment (Orne, 1962). Experiments with financial incentives contingent on participants' decisions minimize this problem because decisions have genuine consequences, as opposed to participation in exchange for a fixed payment or course credit (Hertwig & Ortmann, 2001). In any case, it is worth addressing this concern very carefully as our experiment is unusual in this regard.

Two points must be kept firmly in mind. First, the only mechanism by which experimenter demand can cause differences is by virtue of differences among Treatment conditions. Second, participants in the Anonymous conditions knew that the experimenters would be collecting the data from the envelopes in the bin. Thus, Treatment conditions did not differ insofar as participants expected that the experimenters would eventually know people's choices, whether individually or in aggregate.

One possibility is that the instructions in the nonanonymous conditions caused participants to be concerned about appearing appropriately punitive, causing them to punish more. If so, demand characteristics are not an alternative explanation because this was the point of manipulation. Our interest was in the effect of concern for what others know about one's behavior in the context of moralistic punishment.

If we suppose the operation of the traditional construal of experimenter demand—that participants were motivated to generate data that conform to the predicted effect—then we must ask a great deal of our participants. Because it was a between-participants design, participants would have to: (a) correctly guess what was being varied across conditions; (b) xcorrectly guess how much people in the other condition punished; (c) correctly guess our directional prediction; and (d) choose to punish an amount that conformed to (a)–(c), ignoring other motives (financial or reputational). While this is not impossible, concern for one's reputation is much more plausible.

7.3. Future directions

These results lead to a number of questions to be addressed in future research. First, what specific reputational benefits are gained by being perceived as a third-party punisher? By analyzing people's judgments of punishers and nonpunishers, we hope to understand the reputational gains from moralistic punishment. Second, arguments regarding the putatively modular system underlying punishment suggest that mere cues of social presence, such as eyespots, might exert effects similar to those of actual social presence (e.g., Haley & Fessler, 2005). Determining the conditions that elicit greater punishment can provide insight into the nature of the inputs that activate this computational system.

Other important routes of investigation include: (a) determining the role of intentions, which will help to shed light on models based on avoiding inequities (e.g., Fehr & Schmidt, 1999); (b) determining the role of emotions, which are receiving increasing attention in economic decision making (Fehr & Gächter, 2002; Frank, 1988); and (c) determining the specificity of the effect observed in these experiments (Do similar effects occur in the context of other norm violations, or is there something special about the interactions investigated here?). These lines of research should help illuminate the cognitive adaptations responsible for moralistic punishment.

Acknowledgments

We would like to acknowledge R. Erik Malmgren-Samuel, Erika Tusen, and Alexis Weissberger for contributions to this research project. We would also like to acknowledge helpful comments from participants in the PLEEP.

References

Q5 Barclay, P. (in press). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*.

Barrett, H. C., & Kurzban, R. (in press). Modularity in cognition: Framing the debate. *Psychological Review*.

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

Bolton, G., Katok, E., & Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory*, *27*, 269–299.

Bolton, G., & Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic Behavior*, *10*, 95–121.

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

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

Brown, D. E. (1991). Human universals. New York: McGraw-Hill.

Q5 Burnham, T. C., & Hare, B. (in press). Engineering cooperation: Does involuntary neural activation increase public goods contributions? *Human Nature*.

Butler, J. L., & Baumeister, R. F. (1998). The trouble with friendly faces: Skilled performance with a supportive audience. *Journal of Personality and Social Psychology*, *75*, 1213–1230.

Camerer, C. F. (2003). Behavioral game theory: Experiments in strategic interaction. Princeton, NJ: Princeton University Press.

Q5 Carpenter, J. P. (in press). The demand for punishment. *Journal of Economic Behavior and Organization*.

Carpenter, J. P., & Matthews, P. H. (2004). *Social reciprocity*. Unpublished manuscript.

Carpenter, J. P., & Matthews, P. H. (2005). *Norm enforcement: Anger, indignation or reciprocity*. Unpublished manuscript.

Cottrell, N. B., Wack, D. L., Sekerak, G. J., & Rittle, R. H. (1968). Social facilitation of dominant responses by the presence of an audience and the mere presence of others. *Journal of Personality and Social Psychology*, *9* (3), 245–250.

Cox, J. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*, 260–281.

Cox, J., & Deck, C. (2005). On the nature of reciprocal motives. *Economic Inquiry*, *43*, 623–635.

de Waal, F. B. M. (1996). Good natured: The origins of right and wrong in humans and other animals. Cambridge, MA: Harvard University Press.

Diener, E., Fraser, S. C., Beaman, A. L., & Kelem, R. T. (1976). Effects of deindividuation variables on stealing among Halloween trick-or-treaters. *Journal of Personality and Social Psychology*, *33*, 178–183.

Elster, J. (1998). Emotions and economic theory. *Journal of Economic Literature*, *36*, 47–74.

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

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

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

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.

Fessler, D. M. T., & Haley, K. J. (2003). The strategy of affect: Emotions in human cooperation. In P. Hammerstein (Ed.), *The genetic and cultural evolution of cooperation* (pp. 7–36). Cambridge, MA: MIT Press.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*, 347–369.

Frank, R. (1988). Passions within reason: The strategic role of the emotions. New York: W. W. Norton & Co.

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

Gintis, H. (2005). Behavioral game theory and contemporary economic theory. *Analyse & Kritik*, *27*, 6–47.

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*, 103–119.

Gobin, B., Billen, J., & Peeters, C. (1999). Policing behaviour towards virgin egg layers in a polygynous ponerine ant. *Animal Behaviour*, *58*, 1117–1122.

Q5 Hagen, E. H., & Hammerstein, P. (in press). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology*.

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*, 245–256.

Hauser, M. D., & Marler, P. (1993). Food-associated calls in rhesus macaques (*Macaca mulatta*): II. Costs and benefits of call production and suppression. *Behavioral Ecology*, *4*, 206–212.

Hertwig, R., & Ortmann, R. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*, 383–451.

Hirstein, W. (2005). Brain fiction: Self-deception and the riddle of confabulation. Cambridge, MA: The MIT Press.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. L. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, *7*, 346–380.

Johnstone, R. A., & Bshary, R. (2004). Evolution of spite through indirect reciprocity. *Proceedings of the Royal Society of London. Series B*, *271*, 1917–1922.

Jones, E. E., & Pittman, T. S. (1982). Towards a general theory of strategic self-presentation. In J. Suls (Ed.), *Psychological perspectives on the self*, *Vol. 1*. (pp. 231–262). Hillsdale, NJ: Erlbaum.

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness and the assumptions of economics. *Journal of Business*, *59*, S285–S300.

Ketelaar, T., & Au, W. T. (2003). The effects of guilty feelings on the behavior of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition & Emotion*, *17*, 429–453.

Kurzban, R. (1998). *The social psychophysics of cooperation in groups*. Unpublished doctoral dissertation, University of California Santa Barbara.

Kurzban, R. (2001). The social psychophysics of cooperation: Nonverbal communication in a public goods game. *Journal of Nonverbal Behavior*, *25*, 241–259.

Q5 Kurzban, R., & Aktipis, C. A. (in press). Modular minds, multiple motives. In M. Schaller, J. Simpson, & D. Kenrick (Eds.), *Evolution and social psychology*. New York: Psychology Press.

Latane, B. (1970). Field studies of altruistic compliance. *Representative Research in Social Psychology*, *1*, 49–61.

Ledyard, J. (1995). Public goods: A survey of experimental research. In J. Kagel, & A. Roth (Eds.), *Handbook of experimental economics* (pp. 11–194). Princeton, NJ: Princeton University Press.

Markus, H. (1978). The effect of mere presence on social facilitation: An unobtrusive test. *Journal of Experimental Social Psychology*, *14*, 389–397.

McCabe, K., Smith, V. L., & LePore, M. (2000). Intentionality detection and 'mindreading': Why does game form matter? *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 4404–4409.

Miller, G. F. (2000). The mating mind: How sexual choice shaped the evolution of human nature. New York: Doubleday.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

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

Ottone, S. (2004). *Transfers and altruistic punishments in third party punishment game experiments*. Unpublished manuscript.

Schotter, A., Weiss, A., & Zapater, I. (1996). Fairness and survival in ultimatum and dictatorship games. *Journal of Economic Behavior and Organization*, *31*, 37–56.

Sober, E., & Wilson, D. S. (1998). Unto others: The evolution and psychology of unselfish behavior. Cambridge, MA: Harvard University Press.

Triplett, N. (1898). The dynamogenic factors of pacemaking and competition. *American Journal of Psychology*, *9*, 507–533.

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

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

Wilson, J. Q. (1993). The moral sense. New York: Free Press.

Wright, R. (1995). The moral animal: The new science of evolutionary psychology. New York: Vintage Books.

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*, 205–213.

Zajonc, R. B. (1965). Social facilitation. *Science*, *149*, 269–274.