ELSEVIER

Evolution and Human Behavior 27 (2006) 325–344

Evolution and Human Behavior

Reputational benefits for altruistic punishment

Pat Barclay*

Department of Psychology, Neuroscience, and Behaviour, McMaster University, Hamilton, Canada

Initial receipt 13 September 2005; final revision received 18 January 2006

Abstract

Many studies show that people act cooperatively and are willing to punish free riders (i.e., people who are less cooperative than others). However, nonpunishers benefit when free riders are punished, making punishment a group-beneficial act. Presented here are four studies investigating whether punishers gain social benefits from punishing. Undergraduate participants played public goods games (PGGs) (cooperative group games involving money) in which there were free riders, and in which they were given the opportunity to impose monetary penalties on free riders. Participants rated punishers as being more trustworthy, group focused, and worthy of respect than nonpunishers. In dyadic trust games following PGGs, punishers did not receive monetary benefits from punishing free riders in a single-round PGG, but did benefit monetarily from punishing free riders in iterated PGGs. Punishment that was not directed at free riders brought no monetary benefits, suggesting that people distinguish between justified and unjustified punishment and only respond to punishment with enhanced trust when the punishment is justified.

Keywords: Cooperation; Altruism; Punishment; Public goods games; Trust; Reputation

* Current contact information: Department of Neurobiology and Behavior, Cornell University, Ithaca, NY, 14853, USA. Tel.: +1 607 254 4313; fax: +1 607 254 1303.

E-mail address: pjb46@cornell.edu.

1090-5138/06/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.evolhumbehav.2006.01.003

1. Introduction

In order for altruism among unrelated individuals to evolve, altruists must be able to identify nonaltruists and defectors (Cosmides & Tooby, 1992; Trivers, 1971), and either punish them or avoid them (see, e.g., Axelrod, 1984). This is difficult when altruism cannot be directed toward specific individuals, such as in the provision of public goods. Public goods are things that people have to expend time, effort, or money to provide, but once they are provided, others cannot readily be excluded from benefiting even if they did not contribute the provision of the public good (Davis & Holt, 1993; Messick & Brewer, 1983). Classic examples include irrigation, group protection and vigilance, or any collective action project. Public goods are collectively beneficial, but free riders who cooperate less than fellow group members are better off than people who are more cooperative, causing selection for noncooperation that should eventually undermine collective action. Consistent with this, laboratory subjects reduce their cooperation if others contribute less than themselves to public goods provision (e.g., Andreoni, 1995; Fehr & Fischbacher, 2004a), presumably in retaliation against free riders (Gintis, 2000).

The opportunity to impose sanctions on free riders can potentially solve this collective action problem and allow for the evolution of cooperation because being punished induces free riders to cooperate more (e.g., Boyd & Richerson, 1992; Caldwell, 1976). In non-laboratory settings, such sanctions can include criticism, ostracism, and physical or social threats, all of which carry risks of retaliation, enmity, or loss of partnership. In typical laboratory experiments, the punisher has to pay a monetary cost to reduce the payoff of other players. Despite these costs, some people will punish free riders when they have that option in laboratory experiments (e.g., Fehr & Gächter, 2002; Ostrom, Walker & Gardner, 1992; Yamagishi, 1986) and in field settings (Barr, 2001; Cordell & McKean, 1992; Price, 2005), and this raises the levels of cooperation.

In cooperative group situations, punishing a free rider can be considered a cooperative act because all group members benefit from the resulting increase in the free rider's level of cooperation (Yamagishi, 1986). People without punitive sentiments might be expected to benefit from punishment opportunities more than people who have punitive sentiments and act on them because the former do not pay the cost of imposing sanctions and yet still benefit from the punishment provided by the latter. If this occurred in ancestral environments, then punitive sentiments should have been selected against. Punishment could also decrease in frequency within an individual's lifetime if he/she learns (from experience and observation of others) that punishing brings fewer relative gains than not punishing. People should notice and care that nonpunishers are better off than punishers given that humans care about their payoffs relative to others (e.g., Bolton & Ockenfels, 2000; Roth, 1995), are sensitive to people taking benefits without paying the appropriate costs (Cosmides & Tooby, 1992), and can learn by observation (Tomasello, Kruger & Ratner, 1993). Thus, punishment should decrease in frequency both within generations (via learning) and across generations (as punitive sentiments are selected against), unless there are some processes by which punishment itself is rewarded.

Some game theoretic models and computer simulations of the evolution of punishment postulate that punishers benefit from being in cooperative groups. Groups with sanctions will

have higher cooperation than groups without, so the former will tend to outcompete the latter and will be less likely to disband. The between-group advantage of having sanctions in a group (and consequently higher cooperation) can be greater than the within-group disadvantage that punishers face, so that the level of altruism and altruistic punishment will tend to increase in the population (Boyd, Gintis, Bowles, & Richerson, 2003; Gintis, 2000; Sober & Wilson, 1998). Once punishers become common, there is less need to punish because free riding will be rare, so there is not a big difference between the payoffs to punishers and nonpunishers. Punishment can then be maintained in a group by a weak tendency to imitate the behavior of others (conformist transmission, Henrich & Boyd, 2001). Punishment (and other group-beneficial norms) can spread between populations when less successful groups imitate the norms of cooperative yet punitive (and hence, more successful) groups (Boyd & Richerson, 2002). However, these models are unclear on how punishment becomes common within groups in the first place if punishers are disadvantaged relative to nonpunishers and nonpunishment is the socially prescribed (and hence, most common) behavior. Furthermore, some of these models rely on second-order punishment (punishment of nonpunishers), which has yet to be demonstrated empirically.

If punishers receive personal benefits for their punitive behavior that other group members do not gain, then people could learn to punish. If this also occurred in ancestral environments, then natural selection could have favored the punitive sentiments that motivate punishment (e.g., Fessler & Haley, 2003; Gintis, Smith, & Bowles, 2001; Price, 2003). When punishment is group beneficial, punishers may receive the same type of reputational benefits that altruists receive for their altruism, such as rewards from others. Laboratory experiments (Milinski, Semmann, & Krambeck, 2002a, b; Wedekind & Milinski, 2000) and field research (Gurven, Allen-Arave, Hill, & Hurtado, 2000) both show that more cooperative people receive more rewards from group members than less cooperative people do. It remains to be seen whether people will reward punishers. Altruism could also signal trustworthiness, in that altruists are expected to be less likely to cheat in cooperative partnerships (Alexander, 1987). Barclay (2004) found that people who made high contributions in a cooperative group game were trusted with more money in a subsequent dyadic trust game than those who made lower contributions. When punishing a free rider is good for a group, it could signal the punisher's trustworthiness, commitment to that group, concern with fairness, or unwillingness to tolerate being cheated, such that people trust punishers more than nonpunishers (Fessler & Haley, 2003). This signal need not be a conscious one; it can function as a signal as long as people respond in certain ways to those who display punitive sentiments.

If punishment is a signal of trustworthiness or fairness, for example, then punishers may receive benefits from others who are acting solely out of self-interest. Others might be more willing to enter and invest more in relationships with people who have demonstrated that they will not tolerate unfairness, such that punishers receive more benefits from cooperative partnerships than nonpunishers. Being known for imposing sanctions could be beneficial if other people are less likely to cheat on sanctioners out of fear of retaliation (Brandt, Hauert & Sigmund, 2003). Although punishing nonpunishers and rewarding punishers are altruistic acts that would require explanation themselves (Henrich & Boyd, 2001), trusting and fearing

punishers are not subject to this "second-order" sanctioning problem. It can be in an observer's best interest to enter cooperative relationships with punishers in order to gain a trustworthy partner and avoid cheating in those relationships in order to avoid sanctions. Thus, if there are reputational benefits for punishing, trust and respect (or fear) are likely candidates.

The present set of studies tests the hypothesis that punishers receive reputational benefits for sanctioning free riders. Currently, there are no empirical studies bearing on this hypothesis. The alternative hypotheses are that punishers acquire a bad reputation because of the negative nature of sanctions or that punishing does not lead to reputational consequences. Study 1 examined people's attitudes toward people who punished free riders, and Studies 2–4 tested whether punishers actually received more monetary benefits in experimental trust games than nonpunishers.

2. General methods for public goods game

Undergraduate participants from McMaster University were recruited from an introductory psychology course (in exchange for course credit) and played a cooperative group game known as a public goods game (PGG) with punishment (for details, see Fehr & Gächter, 2002) in groups of four. Each participant was given a pseudonym so that he/she could acquire a reputation in the game yet still remain anonymous, and dividers prevented visual contact between participants. Participants earned "lab dollars" which were exchanged for Canadian dollars after the experiment at a rate of 10 to 1. In each round of the PGG, participants received 10 lab dollars, and they could keep this money for themselves or contribute any amount to a group fund (the public good). The experimenter multiplied the total contributions to the group fund by 1.6 before dividing this new total evenly among the four participants. Thus, contributing was individually costly, yet beneficial for the group, like a prisoner's dilemma with multiple players. After contributing, participants found out what each other participant had contributed and kept, and had the option of paying some of their earnings to punish other participants (of their choice) by reducing those persons' payoffs. Every dollar spent on punishment would reduce the punishee's payoff by $3, and all players were informed after each round of who had punished whom in that round. After the punishment option, a new round began. Participants were forewarned about all aspects of the experimental games (e.g., the presence of punishment and the trust game in Studies 2–4), except that they were never told how many PGG rounds to expect.

3. Study 1 (pilot)

Study 1 gave people experience in PGG with a conspicuous free rider and had them give their views of people who punish free riders and of people who do not. Because of the negative nature of sanctions, punishers will not necessarily be liked more than nonpunishers. However, if punishment signals prosocial qualities like trustworthiness or commitment to a group (Fessler & Haley, 2003), then punishers should be deemed more trustworthy, group

focused, and worthy of respect than nonpunishers. This study also tested whether people's ratings of punishers (relative to nonpunishers) were related to their own punitive behavior.

3.1. Study 1 methods

Thirty male (average age, 18.5±1.1 years) and 22 female (average age, 18.9±0.9 years) undergraduate students played PGG with punishment for five rounds. Instead of playing against other participants, players unknowingly played against computer players who behaved selfishly (contributed \$1, \$1, \$0.5, \$0.5, and \$0 in the five rounds), cooperatively (contributed \$9, \$9, \$8, \$7, and \$7), or relatively neutrally (\$5, \$4.5, \$4, 4, and \$4.5) and responded to punishment by increasing their contributions in the next round by \$1 dollar for every dollar spent to punish them. Participants received punishment for low contributions (\$3) but were not told "who" did the punishing (in this study only). This PGG was part of an experiment that measured PGG punishment after attempting (apparently unsuccessfully) to manipulate participant status. As the manipulated status had no effect on any measures taken (see Barclay, 2005 for details), it will not be mentioned further.

After the PGG, participants were asked to rate hypothetical people who punished or did not punish noncontributors using seven-point Likert scales with anchors of mean/nice, untrustworthy/trustworthy, self-focused/group focused, and unworthy/worthy of respect. The data were analyzed using a repeated-measures general linear model on SPSS (version 11.0), comparing feelings about punishers to feelings about nonpunishers and participant's sex.

3.2. Study 1 results

At least 25% of participants punished in each round, and 88% of participants punished at least once. Men spent more on sanctions than did women [means, \$1.37/round vs. \$0.86; $F(1,50)=4.02, p=.05$]. The majority (87%, 123/141) of the punishment decisions were directed at computer players who contributed less than the punishing participant. Participants who received punishment sometimes (13/141 punishment instances, 9%) lashed out at other players, especially at cooperators. Five instances of punishment (4%) had no obvious provocation.

Participants did not perceive the punishers as being significantly nicer than the nonpunishers ($F<1$), but they did feel that the punishers were more trustworthy, group focused, and worthy of respect [F's(1,48)=7.47, 13.80, 15.13, respectively, all p's<.01] (Fig. 1). There was no interaction of participant's sex with ratings on any of these characteristics [F's(1,48)=2.15, 1.01, 1.89, and 0.02, respectively, all n.s.], nor did sex have any main effects or interactions with other variables (all F's<2.5). Participants who were punished in at least one round for low contributions ($n=38$) did not differ from participants who received no punishment ($n=14$) to the extent to which they rated punishers differently from nonpunishers on any characteristic (all F's<1.2).

There was a significant correlation between how much a participant spent on sanctions and the extent to which he or she thought punishers were nicer than nonpunishers [$r(50)=.36$, $p=.009$]. However, this correlation was not significant for trustworthiness, group focus, or

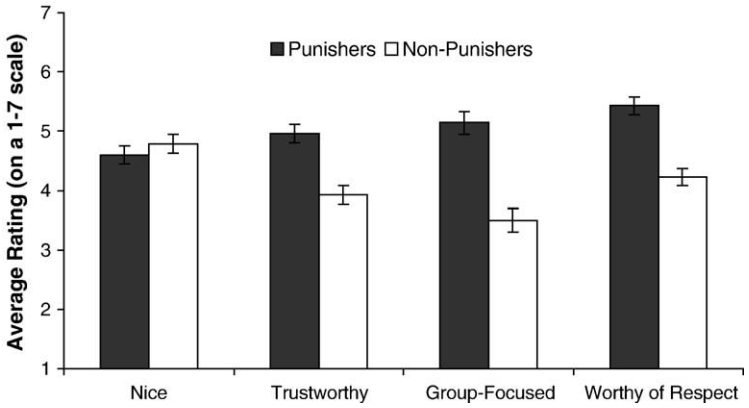

Fig. 1. Average ratings on a seven-point Likert scale of feelings toward punishers (black bars) and nonpunishers (white bars). Higher values represent more positive impressions.

worthiness of respect [*r*'s(50)=.16, .22, .21, *p*'s=.27, .12, .14, respectively]. This suggests that people's attitudes about punishers' trustworthiness, group focus, or worthiness of respect are not simply by-products of the amounts that they spent on punishment, nor were these attitudes significantly predicted by individual contributions [*r*'s(50)=.13, .24, .17, *p*'s=.36, .09, .41, respectively].

3.3. Study 1 discussion

This study showed that after encountering a free rider in PGG, people perceive punishers as being more trustworthy, group focused, and worthy of respect than nonpunishers. This was not due to a general positive impression of the punishers (a "halo effect") because punishers were not seen as nicer than nonpunishers. These perceptions were not affected by participant's sex, nor by whether the participant received sanctions. Also, these results do not appear to have been simply caused by punishers favoring other punishers because there was no significant correlation between individual punishing behavior and the extent to which subjects thought punishers were more trustworthy, group focused, and worthy of respect than nonpunishers were.

4. Study 2

The results of Study 1 were suggestive and could translate to benefits for the punishers if these views affect people's behavior in nonlaboratory environments. However, we need data on whether people will actually invest real money to trust or respect or reward those who apply sanctions. Study 2 tested this by having participants play PGG and then play one round of an experimental trust game with punishers and nonpunishers to see whether they would trust punishers more than nonpunishers.

4.1. Study 2 methods

Twenty-two undergraduates (20 female, 2 male; average age, 18.9±0.9 years) played one round of PGG (with punishment). Instead of real partners, participants unknowingly played against three preprogrammed computer players: a free rider (contributed $1), a punisher (contributed $7, spent $2 to punish the free rider), and a nonpunisher (contributed $7, did not punish).

After the PGG, participants played a modified version of Berg, Dickhaut and McCabe's (1995) trust game using the same pseudonyms from the PGG. In the trust game, players were paired and each received $10. One member of the pair (the truster) could send any number of these dollars to the other member (the responder), and any amount sent got tripled. The responder could then return as much or as little of the tripled amount as he/she desired. Thus, trusters could have increased their payoffs if they trusted the responders, and those responders repaid that trust. To gather data on trust toward each of the other "players," this study used the "strategy method" (e.g., Fehr & Fischbacher, 2004b): participants indicated how much they wanted to entrust to each other player, and these decisions were elicited in random order. Participants were told that those decisions were binding because they would be randomly paired and assigned to roles within a pair, and their corresponding trust decision would be implemented if they were assigned to be the truster in their pair. Thus, the amount entrusted to different "players" in the trust game was a within-subject factor that was analyzed with a repeated-measures general linear model (SPSS version 11.0). The "strategy method" was not used to measure responder behavior. Instead, all participants were told that they had been assigned to be the responder in their pair and which other "player" they were paired with (5 with the free rider, 6 with the nonpunisher, 10 with the punisher, and 1 for whom the data on responder behavior are not available because of a computer error). They were then told that they had been sent $8 and were asked how much of the tripled amount ($24) they wanted to return to the other "player."

4.2. Study 2 results

Participants contributed an average of $5.36 (S.D.=2.82) in the PGG. Of the 22 participants, 14 (64%) punished one of the other "players" (the free rider in 13/14 cases), and the average amount spent on sanctions was $0.91 (S.D.=0.81). Participants' PGG contributions were positively correlated with the amounts they trusted the other three "players" in the trust game on average [$r(20)=.47$, $p=.026$]. However, there was no correlation between participants' punishment and their PGG contributions [$r(20)=.04$, n.s.] or trust in the trust game [$r(20)=.05$, n.s.].

There were significant differences between the amounts entrusted to free riders, punishers, and nonpunishers [$F(2,42)=26.79$, $p<.001$, see Fig. 2], and orthogonal contrasts revealed that free riders were trusted less than contributors (punishers and nonpunishers) [$F(1,21)=27.32$, $p<.001$]. However, punishers and nonpunishers were not entrusted with different amounts ($F<1$). In fact, 15 of the 22 participants sent exactly the same amount to the punisher and the nonpunisher, and of the seven who sent different amounts, four sent

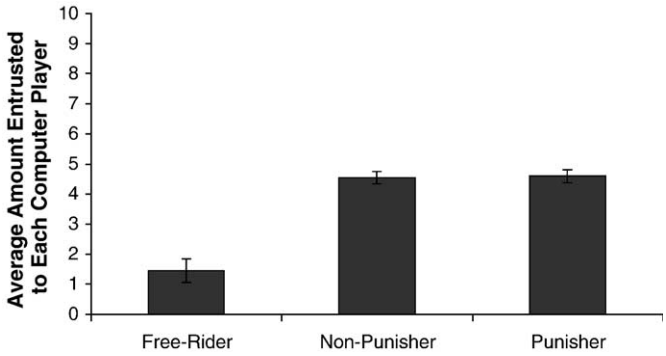

Fig. 2. Average amounts entrusted to free riders, nonpunishers, and punishers in the trust game after one round of PGG in Study 2. Free riders received less than punishers and nonpunishers, yet there were no differences between punishers and nonpunishers.

more to the punisher and three sent more to the nonpunisher. There was no relationship between the amounts that participants contributed or punished in the PGG and how much they entrusted to punishers rather than nonpunishers. There were not enough males to examine sex differences, nor were there enough data points to analyze responder behavior because the "strategy method" was not used.

4.3. Study 2 discussion

Although free riders were trusted less than the contributors (replicating Barclay, 2004), punishers were not trusted more than nonpunishers in Study 2 despite being rated more trustworthy in Study 1. It is possible that because there was only one round, participants did not yet have expertise in the game or strong emotional responses to the actions of others, such as anger toward a repeated free rider. One round does not give players enough information to determine whether a free rider mistakenly made a low contribution or whether he/she will continue to contribute very little (which is more deserving of punishment). Also, there was no chance for participants to see whether the sanctions did induce the free riders to cooperate, so participants may not have realized that punishment of free riders is beneficial to the group, and thus would not have felt gratitude or trust toward punishers. Study 3 addressed these possibilities.

5. Study 3

Study 3 tested whether participants would preferentially trust punishers after repeated exposure to a free rider. It also examined whether participants' own punitive behavior was related to their tendencies to trust (or distrust) punishers. Participants in Study 3 played five rounds of PGG against a strong free rider, a punisher, and a nonpunisher, and then played the same one-shot trust game as in Study 2. To make sure that the reputational effects of contributions and punishment were not confounded with each other, the punisher and

nonpunisher computer players were designed to contribute the same amounts in the PGG but in slightly different orders (for realism), so only the sanctioning behavior would differ between them. The free rider's PGG contributions increased toward the end to simulate the effects of being punished. To reduce the likelihood of the punisher being perceived as malicious or sadistic, the punisher computer player was designed to wait a round before sanctioning and to stop after the free rider started contributing. These computer players were designed to imitate the behavior of freely interacting people.

Having multiple PGG rounds allows for the possibility of "second-order punishing" (i.e., punishing nonpunishers), which several theorists suggest is a significant force in maintaining the existence of punishment and cooperation (e.g., Bendor & Swistak, 2001; Henrich & Boyd, 2001; Sober & Wilson, 1998). If people do engage in second-order punishing, then one would expect that more sanctions would be directed at the nonpunishing computer player than the one that punishes.

5.1. Study 3 methods

Fourteen females (mean age, 18.9 years, S.D.=1.4 years) and 13 males (mean age, 19.0 years, S.D.=1.1 years) played five rounds of PGG with punishment against pre-programmed computer players that they thought were real players. The free-riding "player" contributed \$1, \$0, \$1, \$3, and \$6, respectively, in the five rounds. The two cooperative computer players contributed the same total amount in the PGG, but one (henceforth, "the punisher") punished the free rider by \$0, \$2, \$2, \$3, and \$1, respectively, in the five rounds, whereas the other (henceforth, "the nonpunisher") never punished. Half of the participants saw the punisher contribute \$7, \$8, \$7, \$6, and \$8 in the five rounds and the nonpunisher contribute \$7, \$7, \$6, \$8, and \$8; the contributions of the punisher and the nonpunisher were switched for the other participants. The computer "players" did not change their behavior in response to participants' behavior.

After the PGG, participants played the same modified version of Berg et al.'s (1995) trust game described in Study 2. A median split was used to categorize participants as high or low punishers based on the amounts that they spent on free rider punishment, and this between-subjects categorical variable was added to the within-subject analysis of trust toward free riders, nonpunishers, and punishers. After making trusting decisions, 11 participants were assigned to be responders to the free rider, 8 became responders to the nonpunisher, and 8 become responders to the punisher. All participants were told that their partner had sent them \$8 and were asked how much of the tripled amount (\$24) they wished to return.

5.2. Study 3 results

5.2.1. Public goods game

In the five rounds, participants contributed an average of \$6.3, \$5.5, \$5.6, \$5.6, and \$5.8, respectively. No participants contributed less than the free rider in the three rounds when the free rider's contributions were lowest, so the free rider was a low contributor relative to all participants. As in other studies that include punishment in PGGs, participants' contributions

did not significantly change or decrease across rounds ($F<1$). There was no sex difference in contributions ($F<1$).

Across the five rounds, participants spent an average of \$0.6, \$0.9, \$1.2, \$0.7, and \$0.1 on punishment. There was a significant linear [$F(1,26)=4.74$, $p=.039$] and quadratic [$F(1,26)=12.03$, $p=.002$] component to this pattern, suggesting that participants increased their punishment until the free rider started contributing more because punishment started decreasing in the same round that the free rider started contributing more (round 4). Of the 27 participants, 20 punished at least once. Fifteen of those 20 punished the free rider exclusively, three punished all three other "players" in one or two rounds, one mostly punished the free rider but also punished the nonpunisher in one round, and another exclusively punished the punisher in one round. Participants punished the free rider more than the nonpunisher and punisher [means, \$3.1 vs. \$0.3 and \$0.2, respectively; orthogonal contrast, $F(1,26)=26.43$, $p<.001$] and did not differentially punish the latter two ($F<1$). Although men spent more on punishment than women (means, \$4.5 vs. \$2.7), this difference did not reach significance [$F(1,25)=2.32$, $p=.14$].

5.2.2. Trust game

There were significant differences between the amounts entrusted to free riders, punishers, and nonpunishers [$F(2,50)=43.36$, $p<.001$, see Fig. 3]. An analysis of orthogonal contrasts revealed that free riders were trusted less than punishers and nonpunishers [$F(1,25)=45.55$, $p<.001$], and punishers were trusted significantly more than nonpunishers [$F(1,25)=4.34$, $p=.048$]. Eleven participants entrusted different amounts to the punisher and nonpunisher, and 9 of these 11 trusted the punisher more (binomial test, $p=.033$). There was no overall interaction between participants' own punishment and the amounts they entrusted to all three other "players" [$F(2,50)=1.98$, $p=.15$]. However, the orthogonal contrast was significant [$F(1,25)=6.69$, $p=.016$] for the interaction between participants' own punishment and the amounts entrusted to nonpunishers vs. punishers. An analysis of this interaction reveals that lower-than-average punishers did not entrust different amounts to punishers and

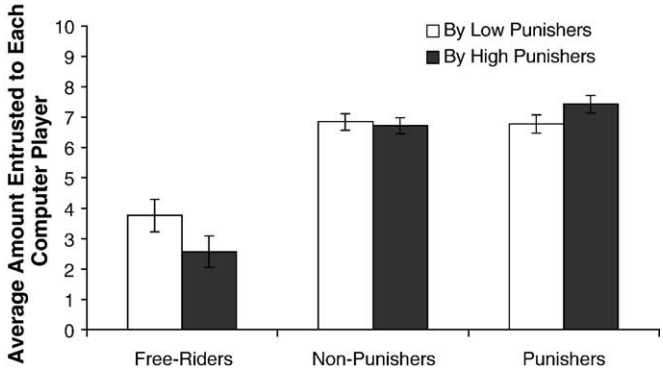

Fig. 3. Average amounts entrusted to free riders, nonpunishers, and punishers in the trust game after five rounds of PGG in Study 3 by participants who provided more (black bars) or less (white bars) than the median amount of punishment. Free riders received less than cooperators, and punishers received more than nonpunishers.

nonpunishers ($F<1$), but higher-than-average punishers trusted punishers more than nonpunishers [$F(1,13)=8.55$, $p=.012$]. Including participant's sex did not affect any of these results.

Participants' PGG total contributions were significantly correlated with the amounts they entrusted to others in the trust game [$r(25)=.57$, $p=.002$]. The amounts that participants returned were significantly correlated with the amounts they themselves entrusted [$r(25)=.58$, $p=.001$] but were not correlated with either their total contributions or punishment [r's(25)=.21 and −.28, both n.s.]. Different amounts were returned to different "players" [$F(2,24)=3.58$, $p=.043$]. Contrast analysis revealed that participants returned less money to free riders than to nonpunishers and punishers ($8.1 vs. $12.9 and $10.3, respectively, $p=.03$) but did not return different amounts to punishers and nonpunishers ($p=.18$). Adding a participant's average trusting behavior as a covariate produced the same results.

5.3. Study 3 discussion

This study replicated Study 2 and Barclay (2004) by showing that people trusted the free rider less than the other "players." More importantly, it showed that punishers were trusted more on average than nonpunishers after five rounds of PGG. It is not surprising that most participants trusted the punisher and nonpunisher similarly: in order to unconfound the effects of contributions and punishment on reputation, the computerized punisher and nonpunisher had to be as similar as possible in contributory behavior. In fact, their total contributions were identical and their behavior only differed in the middle rounds, so primacy and recency effects would make them seem very similar in contributions and punishment. Despite this similarity, a significant number of the participants showing differential treatment trusted the punishers more.

Study 3 was identical to Study 2 but had more rounds and thus longer interaction, so this is probably the factor that caused participants to trust punishers in Studies 1 and 3 but not in Study 2. The computerized free rider in Study 3 was arguably more "deserving" of sanctions than in Study 2 because it continued to contribute relatively little and only increased contributions in response to punishment, which many participants spontaneously commented on after the experiment. Also, participants could observe the positive results from sanctions and could note that the punisher did not punish anyone else. Study 3 did not determine which of these factors is most important in repeated interactions but did show the effects of prolonged exposure to free riders and punishers. It is interesting that punishers were trusted more but were not returned more money in the trust game. This suggests that people may trust punishers but do not reward them more than nonpunishers. If so, this would support the idea that people treat punishment as a signal of trustworthiness (despite the fact that participants' punishment was not correlated with the amounts they returned in this experiment).

There was no evidence that participants performed much second-order punishing in this experiment. Participants in this study saw a clear free rider that one "player" failed to punish, but this player did not receive more punishment than the "player" that did punish. Participants punished the nonpunisher as often as the punisher, and there was no difference in the amounts they were punished, contrary to the predictions of models involving second-order punishment (e.g., Bendor & Swistak, 2001; Henrich & Boyd, 2001; Sober & Wilson, 1998). This finding is

consistent with the results of Kiyonari, Shimoma, and Yamagishi (2004) and Kiyonari and Barclay (2005) and weakens any theoretical models that rely on second-order punishment to maintain punishment.

6. Study 4

Study 4 sought to replicate the findings of Study 3 and test whether they would occur in PGG games with naturally occurring variation. Participants in Study 4 played five rounds of PGG and then the one-shot trust game from Studies 2 and 3. Participants experienced naturally occurring variation in cooperation and punishment because computer players were not used in this study. If the effects of punishment on trustworthiness are similar, then this naturally occurring variation allows us to generalize the results a bit more than we could with artificially occurring variation alone, and it allows us to examine different types of punishment. Some natural punishment is arguably justifiable because it is directed at free riders, while some are not because it is directed at cooperators. In Study 4, it was possible to examine the differential effects of justified and unjustified punishment on people's trustworthiness.

6.1. Study 4 methods

6.1.1. Participants and procedure

Fifteen males (average age, 20.6±4.5 years) and 45 females (average age, 18.8±1.0 years) played five rounds of PGG with punishment in 15 groups (of four participants each) without computer players and then played the modified version of Berg et al.'s (1995) trust game described in Study 2. As in Studies 2 and 3, Study 4 gathered data on how much participants would trust each of the other players by asking participants how much money they would entrust to each potential recipient (strategy method), and I randomly formed the pairs afterward. In each pair, the participant who was assigned to be the responder was told what he/she received from the truster and then decided how much of the tripled amount to return to the truster. Thus, the strategy method was not used for responder decisions, so there were far fewer data points for this variable (30 total, each with a different amount sent).

6.1.2. Statistical analysis

Punishment was coded as either justified or unjustified according to a defined algorithm. Justified punishment was defined as sanctions imposed on the lowest contributor (the "free rider") in the group on any given round or punishment from the bottom up if more than one person was punished, as long as the punished person had contributed less than the punisher. This definition was chosen in order to exclude punishment that singled out any nonlowest contributor without also punishing the lowest contributor because the punished person would likely feel such behavior was unjustified and it would be hard to distinguish such punishment from aggression.[1] Unjustified punishment was defined as any other sanctions, which included

[1] Justified punishment is a significant predictor of trust even if it is redefined as "punishment of any lower-than-average contributor," so results will only be presented for the analysis with the original definition.

retaliation or punishment of people who contributed as much or more than the punisher in the given round.

To test whether the participants trusted justified punishers more than players who did not provide justified punishment, we conducted a multiple linear regression analyses (SPSS version 12.0) to see what factors would predict the amount each participant was entrusted with. For each recipient, the regression examined the average amount of money that the other group members were willing to entrust to him/her. Past research has found that there is a correlation between the total PGG contributions in a group and the level of trust displayed by group members in subsequent trust games (Barclay, 2004). Furthermore, groups with strong free riders are likely to require justified punishment, but the low PGG contributions by the free rider will also bring down the group's level of trust Thus, it is important to statistically control for group levels of trust. For this reason, 15 dummy variables were created to indicate membership or nonmembership in each group (one for each group, each with a value of 1 or 0 for each participant). These dummy variables factor out the general levels of trust exhibited by each group in order to compare each person to his/her group, which is the relevant comparison group to test against when testing for a within-group advantage of punishing.

6.2. Study 4 results

6.2.1. Public goods game

In the five rounds, participants contributed an average of \$5.5, \$6.0, \$7.0, \$7.5, and \$7.8, respectively, which represents a significant linear increase [orthogonal contrast, $F(1,14)=25.84$, $p<.001$]. Of the 60 participants, 23 provided no punishment, 19 provided only justified punishment (average, \$2.9 spent), 7 provided only unjustified punishment (average, \$3.6 spent), and 11 provided both (average, \$3.3 and \$3.4 spent, respectively). There were 64 instances of justified punishment and 49 instances of unjustified punishment. Of the latter, 23 were retaliation for punishment in the previous round,[2] 6 were retaliation delayed by one round, 12 were free riders punishing everyone (9 of which occurred while retaliating against someone), 3 were free riders punishing the highest contributor alone,[3] 2 were free riders punishing the second lowest contributors ("hypocritical punishment"),[4] 2 were delayed punishment of free riders, and 1 was a participant apparently retaliating on behalf of someone else. There were no unambiguous cases of people punishing nonpunishers (second-order punishment). Punishment did not change in frequency across rounds ($F<1$).

Men did not contribute more in the PGG than women did (\$36.1 vs. \$32.9, $t=1.10$, $p=.28$), but they spent more on sanctions than did women (\$4.4 vs. \$1.9, $t=2.72$, $p=.009$). Men spent more on justified punishment than women (\$2.7 vs. \$1.1, $t=2.23$, $p=.026$),

[2] Some of these were retaliation by non-free riders against unjustified punishment, which is arguably justifiable. However, recoding this type of retaliation as justified punishment does not affect the results.

[3] Punishment of contributors is often found at nonzero levels (e.g. Fehr & Gächter, 2000) and has been interpreted as generalized or preemptive retaliation.

[4] This was coded as unjustified because hypocrisy is an unjustified behavior virtually by definition. Recoding this "hypocritical punishment" as justified punishment does not affect the results.

although not more on unjustified punishment ($1.7 vs. $0.8, $t=1.27$, $p=.22$). After controlling for group contributions, a participant's PGG contributions were positively correlated with the amount of justified punishment he/she provided ($r=.33$, $p=.011$) and negatively correlated with unjustified punishment ($r=-.36$, $p=.006$).

6.2.2. Trust game

Group trusting behavior was an important determinant of amounts received in the trust game because all of the dummy variables for group membership were significant (average standardized β for dummy variables=.548, all p's<.05). Justified punishment positively predicted the amounts received in the trust game (standardized $\beta=.232$, $t=2.21$, $p=.032$), whereas unjustified punishment negatively predicted the amounts received (standardized $\beta=-.227$, $t=-2.23$, $p=.031$) and contributions in the PGG were not a significant predictor (standardized $\beta=-.019$, $t=-0.130$, $p=.90$). The regression model was highly significant [$F(17,42)=12.54$, $p<.001$] and accounted for 77% of the variance in amounts received. Amounts sent to people in the trust game (A) can be described by the following equation:

$$A = 2.644 - 0.004\ c + 0.196\ j - 0.210\ u + g + \varepsilon \tag{1}$$

where c is the recipient's total contributions, j is his/her justified punishment, u is his/her unjustified punishment, g is the group-specific "effect" (i.e., the group dummy variable and its coefficient; these have an average value of 4.199), and e is unaccounted variance (error).

One potential problem with these analyses is that free riders in a group cannot be justified punishers because the definition of justified punishment precluded participants who punished anyone that contributed more than themselves. Therefore, justified punishers may be entrusted with more money than nonpunishers simply because participants did not entrust money to free riders, and justified punishers could not be free riders. However, recipients' justified punishment was a stronger predictor of amounts received than PGG contributions were, which speaks against this potential criticism. To further examine this potential problem, we rerun the analysis after excluding the lowest contributor in each group, who by definition could not be a justified punisher, leaving 45/60 participants. The regression model was still significant [$F(17,27)=8.35$, $p<.001$] and accounted for 74% of the variance in amounts received. Even with this reduced power, justified punishment predicted amounts received, although this just failed to reach significance (standardized $\beta=.319$, $t=1.98$, $p=.058$). The fact that it was a marginally significant positive predictor of amounts received (even after reducing power by excluding the lowest contributor in each group) suggests that justified punishers were not trusted more merely because they were not free riders. Neither unjustified punishment (standardized $\beta=-.069$, $t=-0.54$, $p=.59$) nor PGG contributions (standardized $\beta=-.154$, $t=-0.65$, $p=.52$) predicted amounts received in this reduced sample. The amounts sent to non-free riders in the trust game (A) can be described by the following equation:

$$A = 5.315 - 0.033\ c + 0.312\ j - 0.084\ u + \sum g + \varepsilon \tag{2}$$

Eq. (2) uses the same symbols as Eq. (1), except that the average group-specific coefficient (g) has a value of 2.232.

The amount sent by a truster significantly predicted the amount returned to him/her in the trust game (standardized β=.447, p=.033). Although the regression model was significant [$F(18,11)$=2.69, p=.049] and accounted for 51.2% of the variance in amounts returned to trusters, no other variables significantly predicted amounts returned to trusters because there were far fewer observations to test so many variables (including group dummy variables); only half of the participants were responders, they only returned money to one truster each, and two trusters did not send anything so there were no data for them or their responders. The only thing that significantly predicted how much money a given responder returned was the amount that the truster sent to him/her (standardized β=.615, p=.002), and this model was significant [$F(18,11)$=3.82, p=.012] and accounted for 63.7% of the variance in responder behavior. Responders returned an average of 46% of the tripled amount that they received.

Some groups earned more money than others; in 10 of the 15 groups, membership in that group was a significant positive predictor of individual earnings (average standardized β for group dummy variables=.337). The only significant individual-level predictors of earnings were a participant's unjustified punishment, which negatively predicted earnings (standardized β=0 to −.276, t=−2.33, p=.025), and whether the subject was the responder in the trust game, which positively predicted earnings (standardized β=.275, t=3.71, p=.001). Although neither justified punishment nor individual contributions significantly predicted earnings once group earnings were accounted for (standardized β's=.104 and −.296, t's=0.85 and −1.75, p's=.40 and .088, respectively), it is important to note that justified punishment was a positive (albeit nonsignificant) predictor of earnings. Thus, the increased trust toward justified punishers did seem to eliminate the payoff disadvantage that they would otherwise experience relative to nonpunishers. This regression model was highly significant [$F(18,41)$=8.32, p<.001] and accounted for 69% of the variation in earnings in the experiment.

6.3. Study 4 discussion

Previous studies have shown that opportunities to punish others and gain a reputation for trustworthiness can cause PGG contributions to increase across rounds (e.g., Barclay, 2004; Fehr & Gächter, 2002), and this study showed that the two effects together cause an increase in contributions. This study also found that men punished more than women. This was not simply aggression because men provided more justified punishment but not more unjustified punishment. Fehr and Gächter (2000) suggested that some punishments of cooperators are retaliation, and this study shows that retaliatory punishment occurs frequently when participants can see who imposed sanctions. Approximately 20% (23/113) of the instances of punishment were apparently in retaliation for other punishment. Such a high frequency suggests that punishment is not costless, as some theorists have argued (e.g., Sober & Wilson, 1998).

This study also replicates Study 3 in failing to find unambiguous evidence for second-order punishing (punishing nonpunishers) despite the existence of multiple rounds. Theorists have predicted that this is an important component of the evolution of cooperation (e.g., Henrich & Boyd, 2001; Sober & Wilson, 1998), but other studies have also noted a conspicuous lack of second-order punishment (Kiyonari & Barclay, 2005; Kiyonari et al.,

2004). If second-order punishment is not common, then other processes (such as increased trustworthiness of punishers) must have supported the evolution of punitive sentiments.

After controlling for the trusting behavior of other group members, justified punishment did significantly predict how much a person was trusted with. Justified punishers were even trusted somewhat more than other cooperators, which suggests that the effect was not simply because justified punishers were trusted more than free riders. Thus, this study replicates Study 3 by showing that people trust more money to those who have demonstrated justified punishment, although neither study provided evidence that people return more money to them. Not all punishment led to trustworthiness because recipients' unjustified punishment negatively predicted how much they were entrusted with. It is somewhat surprising that participant's PGG contributions did not predict the amounts they were trusted with because this was found consistently in Studies 2 and 3 and in past research (Barclay, 2004). People who made high PGG contributions tended to provide more justified and less unjustified punishment, and both of those were significant predictors of amounts received in the trust game. I suspect that those two variables together accounted for the trust that would otherwise have been attributed to PGG contributions, such that contributions did not predict anything beyond that which was predicted by punishment. However, contributions are probably a necessary component of the trustworthiness signal, such that punishment without contributions is seen as hypocrisy and is judged unjustified. Alternate explanations (provided by reviewers) are that the group dummy variables accounted for the trust that would have otherwise been attributed to individual PGG contributions, or that punishment reduces the signaling value of cooperation because even free riders will cooperate in the face of punishment.

This study cannot speak to the question of whether justified punishers actually were more trustworthy than nonpunishers because there were relatively few data points on responder behavior. Justified punishers may be very discriminating in their trustworthiness, such that they repay the trust of cooperators but not of free riders, just as high contributors tend to do (Albert, Güth, Kirchler, & Maciejovsky, 2002). This is especially likely to happen with punishers because the act of punishment demonstrates a dislike of free riders that could easily cause them to repay the trust of free riders less than nonpunishers would. Thus, punishers might not be more trustworthy overall, but only toward cooperators, and future studies could investigate this by varying the level of cooperativeness of punisher's partners.

An informal game theoretical analysis provides another explanation as to why justified punishers did not appear to be especially trustworthy and also predicts why we would not expect punishers to actually earn more with this particular design. Because there was only one round of the trust game, it could never occur that a cooperative signal could be sufficiently costly to deter cheaters yet still pay off to the signaler. If the benefits of being trusted in a one-shot trust game ever did outweigh the cost of cooperation (and punishment), then we would expect to see dishonest signals of cooperation and punishment from people who intended to cheat in the trust game. Given that possibility, observers would be expected to discount signals of cooperative intent to a level where their trust does not completely compensate the signaler for the cost of signaling, and this would deter dishonest signalers. Cooperative signals could pay off if there were multiple interactions: the signal pays off in the long run for an honest signaler who intends to cooperate repeatedly but does not pay off for a dishonest signaler who

intends to defect in the first interaction (which invites mutual defection). Given that this experiment used a one-shot trust game to measure levels of trust (in order to avoid strategies associated with repeated play), one would expect some dishonest signaling and also some discounting of cooperative signals such that the enhanced trust does not completely make up for the cost of cooperation. Thus, it is all the more impressive that justified punishers were apparently partially compensated for the cost of punishment with increased trust, such that they did not earn less than those who provided less punishment (and in fact, earned slightly but not significantly more). Future studies will give and test a more formal model of signaling cooperative intent.

Finally, one may question why people would provide unjustified punishment, seeing as it led to a decrease in earnings. Over three quarters of unjustified punishment involved retaliation of some sort, so it may serve the function of deterring future punishment or aggression, just as other forms of aggression may function to deter future transgressions (e.g., Daly & Wilson, 1988). Thus, unjustified punishment may ultimately benefit punishers in ways that were not tested with the current methodology.

7. General discussion

Study 1 demonstrated that people rated altruistic punishers as more trustworthy, group focused, and worthy of respect than nonpunishers; Studies 3 and 4 supported this by finding that justified punishers were trusted more than nonpunishers and received monetary benefits for punishing. However, Study 2 did not find such an effect. The most likely explanation for the different results is that participants played five rounds of PGG in Studies 1, 3, and 4 but only one round in Study 2. Punishment increased perceived trustworthiness in all the studies where there were repeated interactions (1, 3, and 4) but did not in any study where there was only one round (Study 2, an unpublished replication of Study 2, and the one-round PGGs in Kiyonari & Barclay, 2005, and Kiyonari et al., 2004). Five rounds allow for more time to gain expertise in the game and to experience emotional responses toward free riders and punishers. After only one round, there may be too little information to accurately guess the motivations of the free riders and the punishers and to tell whether punishment of a free rider is truly justified. In Study 4, justified and unjustified punishment had opposite effects on one's reputation, so if the justification for punishment was unclear in Study 2, then those effects would cancel each other out. One might expect trust toward free riders to be less affected by this ambiguity of intentions than trust toward punishers and nonpunishers because the cost of mistaking a selfish free rider from a hesitant cooperator (both low contributors) is likely to be greater than the cost of mistaking a justified punisher for a cooperative thug (both of whom are cooperators who punish). Finally, five rounds allow participants to appreciate the effects of sanctions on free riders, which may be necessary for people to trust punishers. Future studies can test whether punishment needs to be effective in order to bring reputational benefits. Punishments might bring reputational benefits even after single-round PGGs, provided that participants were (a) already familiar with the game, (b) aware that punishment was good for everybody, (c) aware of the intentions behind the punishment, and (d) sufficiently emotionally aroused

toward the free riders to demand their punishment, but these conditions would be difficult to achieve with a single interaction in an unfamiliar experimental situation.

Together, these results suggest that costly sanctioning of free riders might not actually be costly once there are opportunities for the punisher to acquire a reputation. Punishers provide a public good by forcing free riders to cooperate, and people do seem to realize this after playing multiple rounds of PGG. There may be variance in the circumstances that justify punishment, so we might expect reputational benefits to accrue only to people who perform punishment that is considered justified in a given culture. Whether this makes up for the costs of punishment depends on the frequency of collective action projects (and free riders to punish) and dyadic opportunities for trust. If dyadic interactions are more frequent or carry larger potential payoffs than collective action projects, then the reputational benefits of punishing could easily compensate the punishers for more than the cost of the altruistic punishment, such that justified punishers actually do better than nonpunishers. Future studies should test whether punishers are similarly rewarded or trusted outside the laboratory. If so, then reputation could eliminate the disincentive to punish free riders and cause punishment to increase in frequency in populations via individual learning. If such benefits were also accrued in the ancestral environments in which humans evolved, then reputation (with or without group-level effects) could explain why the psychological mechanisms that modulate altruism and altruistic punishment evolved. This argument does not require psychological mechanisms designed specifically for according reputation to punishers; if people do treat punishers better (whether this is "learned" or "innate" or a combination of both), then it can provide a selective pressure for punitive sentiment. Given the present findings, reputational benefits certainly appear to be more important in the evolution of punitive sentiment than is second-order punishment, which was not a factor at all in these studies (Trivers, 1971).

Acknowledgments

The author thank M. Wilson, M. Daly, A. Muller, D. Krupp, A. Clark, L. Debruine, and R. Morrison for comments and advice. P. Ramos wrote the computer programs, and M. Mackenzie helped collect data. The Social Sciences and Humanities Research Council (SSHRC) supported this research with a grant to M. Wilson and a doctoral fellowship to P. Barclay.

References

Albert, M., Güth, W., Kirchler, E., & Maciejovsky, B. (2002). Are we nice(r) to nice(r) people? An experimental analysis Discussion Paper 2002-15. Jena, Germany: Max Planck Institute for Research into Economics Systems, Strategic Interaction Group ftp://papers.mpiew-jena.mpg.de/esi/discussionpapers/2002-15.pdf.

Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine de Gruyter.

Andreoni, J. (1995). Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review*, *85*, 891–904.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons". *Evolution and Human Behavior*, *25*, 209–220.

Barclay, P. (2005). *Reputational benefits of altruism and altruistic punishment.* PhD Thesis. McMaster University.

Barr, A. (2001). Social dilemmas and shame-based sanctions: Experimental results from rural Zimbabwe Working Paper WPS/2001.11. University of Oxford, UK: Centre for the Study of African Economics.

Bendor, J., & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology*, *106*, 1493–1545.

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

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*, 166–193.

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

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

Boyd, R., & Richerson, P. (2002). Group beneficial norms can spread rapidly in a structured population. *Journal of Theoretical Biology*, *215*, 287–296.

Brandt, H., Hauert, C., & Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London Series B*, *270*, 1099–1104.

Caldwell, M. D. (1976). Communication and sex effects in a five-person Prisoner's Dilemma game. *Journal of Personality and Social Psychology*, *33*, 273–280.

Cordell, M. A., & McKean, J. (1992). Sea tenure in Bahia, Brazil. In D. W. Bromley (Ed.), *Making the commons work: Theory, practice, and policy* (pp. 183–205). San Francisco: ICS Press.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Daly, M., & Wilson, M. (1988). *Homicide*. New York, NY: Aldine de Gruyter.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics*. Princeton: Princeton Univ Press.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends Cogn Sci*, *8*, 185–190.

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

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

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

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

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*, 160–179.

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

Gurven, M., Allen-Arave, W., Hill, K., & Hurtado, M. (2000). "It's a wonderful life": Signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior*, *21*, 263–282.

Henrich, H., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*, 79–89.

Kiyonari, T., & Barclay, P. (2005). Selective incentives for cooperation: Second-order punishment vs. second-order reward. Talk presented at the 17th Annual Meeting of the Human Behavior & Evolution Society, Austin, Texas (June 2005).

Kiyonari, T., Shimoma, E., & Yamagishi, T. (2004). Second-order punishment in a one-shot social dilemma. Poster presentation at the 28th International Congress of Psychology, Beijing, China (August 2004).

Messick, D. M., & Brewer, M. B. (1983). Solving social dilemmas: A review. In L. Wheeler, & P. Shaver (Eds.), *Review of personality and social psychology*, *Vol. 4* (pp. 11–44). Beverley Hills, CA: Sage Publications Inc.

Milinski, M., Semmann, D., & Krambeck, H. -J. (2002a). Reputation helps solve the "tragedy of the commons." *Nature*, *415*, 424–426.

Milinski, M., Semmann, D., & Krambeck, H. -J. (2002b). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society of London Series B*, *268*, 881–883.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, *86*, 404–417.

Price, M. (2005). Punitive sentiment among the Shuar and in industrialized societies: Cross-cultural similarities. *Evolution and Human Behavior*, *26*, 279–287.

Price, M. E. (2003). Pro-community altruism and social status in a Shuar village. *Human Nature*, *14*, 191–208.

Roth, A. E. (1995). Bargaining experiments. In J. H. Kagel, & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 253–348). Princeton, NJ: Princeton University Press.

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

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*, 495–552.

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

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850–852.

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