J Evol Econ (2004) 14: ♣–♣
DOI: 10.1007/s00191-004-0212-1

Journal of Evolutionary Economics

Why Punish? Social reciprocity and the enforcement of prosocial norms*

Jeffrey P. Carpenter[1], **Peter Hans Matthews**[1] **and Okomboli Ong'ong'a**[2]

[1] Department of Economics, Middlebury College, Middlebury, VT 05753, USA
(e-mail: {jpc,pmatthew}@middlebury.edu)
[2] Gradute School of Business, Stanford University, Stanford, CA 94305, USA
(e-mail: ongonga@gsb.stanford.edu)

Abstract. Recently economists have become interested in why people who face social dilemmas in the experimental lab use the seemingly incredible threat of punishment to deter free riding. Three theories with evolutionary microfoundations have been developed to explain punishment. We survey these theories and use behavioral data from surveys and experiments to show that the theory called social reciprocity in which people punish norm violators indiscriminately explains punishment best.

Keywords: Social dilemma – Punishment – Norm – Evolutionary game theory – Experiment

JEL Classification: C91, C92, D64, H41

1 Introduction

Economists have become interested in analyzing, in the experimental lab, something known to field researchers for quite a while, that people who face social dilemmas (i.e. situations in which group and individual incentives are at odds) sometimes control free riding locally by the use of social, economic, and/or physical sanctions.[1] The existence of schemes by which people monitor each other and

* We thank Carolyn Craven, Corinna Noelke and two referees for comments, and Middlebury College for financial assistance. In addition, Carpenter acknowledges the support of the National Science Foundation (SES-CAREER 0092953).

Correspondence to: J.P. Carpenter

[1] Economic punishment experiments include .Fehr and Gaechter (2000), .Bochet et al. (2003), .Bowles et al. (2001), .Carpenter (2002b), .Carpenter and Matthews (2002), .Masclet et al. (2003), and .Sefton et al. (2000). Relevant field research is summarized in .Ostrom (1990) and .Ostrom et al. (1994). A specific example of field research is .Acheson (1988).

punish those who free ride is problematic for standard economic theory. *Why?* First, in non-repeated interactions, any theory assuming that agents simply want to maximize their material gain can not reconcile the cooperative behavior needed to obtain socially efficient outcomes because free riders always do better. Free riding when others contribute avoids the costs associated with contributing yet returns the benefits of cooperation and free riding when others free ride prevents one from being taken advantage of. Hence, no self-interested person would ever cooperate. The same argument can be made for not punishing free riders because punishment, in this context, is just a second-order social dilemma (see .()Boyd and Richerson (1992). Those people who don't punish avoid the costs of doing so, but share any benefits associated with the punishment inflicted by others.

Understanding punishment in social dilemma games has become an interdisciplinary endeavor with theories being offered that have economic, evolutionary psychological, and biological foundations. In this paper we describe each theory and discuss data from experiments and surveys to evaluate their behavioral relevance.

We begin, in Section 2, by providing microfoundations for punishing behavior by building a simple model of the evolution of behavior in an institutional environment that permits the monitoring and sanctioning of free riders. From this model we identify two types of punishers: those who punish free riders in their groups and those who punish free riders outside their immediate group. In Section 3 we describe three theories that have been developed to explain why punishment occurs in experiments with structures similar to our stylized model. In Section 4 we describe various sources of data which we use to evaluate each theory. We conclude in Section 5 by discussing the support each theory finds in the data.

2 The evolution of punishing behaviors

For simplicity, imagine a two-person social dilemma modeled on the widely used experimental game, the voluntary contribution mechanism (Isaac et al., 1984), in which agents are given 50 experimental monetary units (EMUs) and allowed to either contribute to a public good or keep the money for themselves.[2] If a player contributes, her money is multiplied by a factor of 1.5 and shared equally with the other group member. If she doesn't contribute, i.e. she free rides, she keeps her 50 EMUs and may receive another 37.5 EMUs if the other person contributes. Clearly, free riding is the dominant strategy.

2.1 Microfoundations for ingroup punishment

Now we introduce punishment using a mechanism that is also consistent with the protocols used in most punishment experiments. At a cost of 1 EMU players can buy a 2 EMU reduction in the other player's payoff. For simplicity we make

[2] Because our purpose is to evaluate different theories of punishment, the model we present here is a shorter, simplified version of the model developed in Carpenter and Matthews (2002).

Table 1. The normal form of the social dilemma with punishment game

	Free ride		Cooperate		Punish	
Free ride	50,	50	87.5,	37.5	67.5,	27.5
Cooperate	37.5,	87.5	75,	75	75,	75
Punish	27.5,	67.5	75,	75	75,	75

both the contribution and punishment decisions binary; players contribute all 50 EMUs or none and they spend either 10 EMUs to punish a free rider or nothing. We also restrict the strategy space to those behavioral types that we actually see with any regularity in the experimental lab. Essentially, this means that, while free riders occasionally punish cooperators, they do so rarely enough that we restrict the ability to punishment to cooperators. This gives us three behavioral types: *Free Riders* who neither contribute nor punish, *Cooperators* who contribute but never punish, and *Punishers* who contribute and punish free riders in their groups. The normal form of this game is presented in Table 1. Each player has two components to her strategy, a contribution decision and, if she contributes and her partner free rides, a punishment decision.[3]

There are two pure strategy Nash equilibria in the normal form of the game: (1) both players free ride and do not punish and (2) both players contribute and punish free riders. However, only one of these equilibria, the first, is subgame perfect in the extensive form of the stage game that is used in the experimental lab. As mentioned above, punishment is a second-order public good meaning players can do better by free riding off any punishment doled out by other players. Further, punishment is an incredible threat because it is costly to engage in and therefore not punishing dominates punishing. Knowing punishment is dominated, free riders should not fear it and therefore the subgame perfect equilibrium is where players free ride and do not punish. However, subgame perfection is a very restrictive refinement. We now examine the implications of evolutionary dynamics that sometimes select other, more intuitive Nash equilibria.

We describe the evolution of behavioral types in this population using the familiar replicator dynamics (Taylor and Jonker, 1978). At each moment in continuous time, nature randomly assigns two people from a large population to play our game. Under this dynamic, behavioral types succeed in the population to the extent that they do better or worse than the average agent. Where $S = \{FR, C, P\}$ is the set of allowed strategies, $p_{i,t}$ is the fraction of players using strategy i in period t, and $\pi_{i,t}$ is the current payoff of using strategy i in the population, the average payoff to all strategies in period t is:

$$\overline{\pi_t} = \sum_{i \in S} p_{i,t} \pi_{i,t}$$

[3] Note this game is a version of "norms game" described in Sethi (1996), which is based on Axelrod (1984). See also Binmore and Samuelson (1994) and Gueth and Kliemt (1993).

According to the replicator dynamics the growth of strategy i is described by:

$$p_{i,t+1} = p_{i,t}\left(\frac{\pi_{i,t}}{\overline{\pi_t}}\right)$$

which expressed as a difference equation is just:

$$p_{i,t+1} - p_{i,t} = p_{i,t}\left(\frac{\pi_{i,t} - \overline{\pi}}{\overline{\pi_t}}\right)$$

Because the denominator of the right side of the equation does not affect the value of the zeros of the dynamic (only the speed to an equilibrium), it is standard to focus on the numerator only. In the limit, as the difference between t and $t+1$ becomes arbitrarily small, the discrete time dynamic collapses to the following continuous time derivative with rest points at all the Nash equilibria of the underlying game and asymptotically stable rest points containing only the evolutionarily stable strategies (ESS) of the game.

$$\dot{p_i} = p_i(\pi_i - \overline{\pi})$$

Letting $\pi(i,j)$ be the payoff of playing strategy i against strategy j, calculating the expected payoffs of each behavioral type from the normal form that corresponds to Table 1 is straight-forward,

$$\pi_i = \sum_{i,j \in S} p_j \pi(i,j)$$

and yields the following results:

$$\begin{aligned}
\pi_{FR} &= 67.5 - 17.5p_{FR} + 20p_C \\
\pi_C &= 75 - 37.5p_{FR} \\
\pi_P &= 75 - 47.5p_{FR}
\end{aligned}$$

after substitution for the residual share $p_P = 1 - p_C - p_{FR}$.

Simulating the dynamic from different initial conditions will give us an idea of whether punishing behavior can survive selection under the structure and payoffs of the typical punishment experiment. Intuition says that punishing strategies can not successfully invade a free riding population because they are much more likely to be matched with a free rider than another punisher in which case they do worse ($\pi_P = 27.5$) than the average ($\overline{\pi} \doteq 50$). Likewise, because cooperators do not share the burden of punishing free riders and therefore cannot reduce their fitness, initial populations with too many cooperators will be vulnerable to invasion by free riders because there are not enough punishers. However, when punishers are significantly represented in the initial population, free riders are likely to be matched with one who reduces their payoffs. Further, the high probability of being punished means the free rider's payoff will be driven below the average which is now disproportionately weighted by interactions between types who cooperate with each other and therefore do not punish each other. After the free riders are driven from the population there

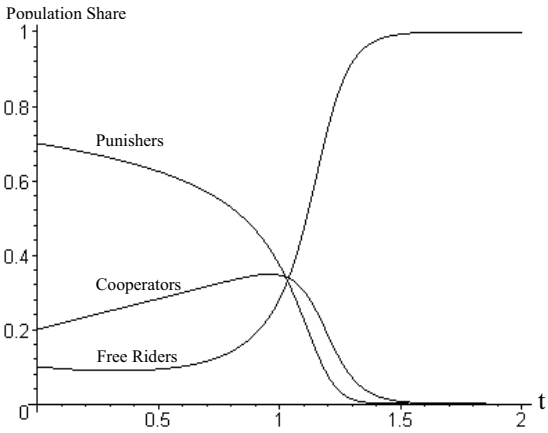

Fig. 1. The normal form of the social dilemma with punishment game

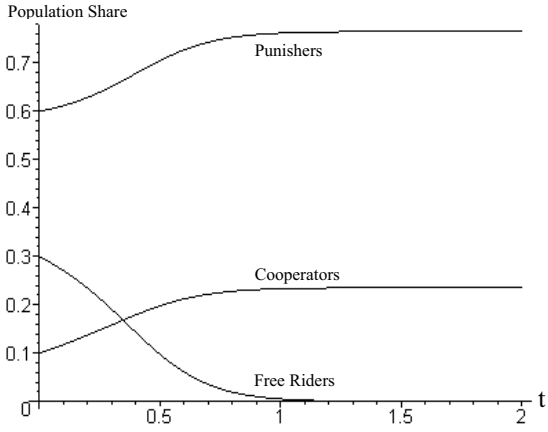

Fig. 2. Punishment and cooperation can not invade a population of free riders

is no selection pressure against punishing types and the population shares stabilize. These dynamics are illustrated in Figures 1 and 2.

As one can see, Figure 1 illustrates the case in which there are relatively too many cooperators in the starting distribution of types to stave off an invasion by free riders. Initially, cooperators increase in the population because they free ride on the punishment doled out by the initial group of punishers, but punishers do poorly relative to cooperators and free riders who interact with cooperators and therefore begin to wane in the population. At some point, the number of punishers falls to a level at which free riding takes off in the population and after a short while the population fixes at the all free riding subgame perfect Nash equilibrium.

There is a second component of (non-perfect) Nash equilibria, however, in which players mix between cooperation with and without punishment such that the probability of the former is at least 62.5%. In terms of our evolutionary model, this corresponds to a polymorphism in which all individuals are contributors and most

are prepared to punish free riding. The component is (at least) a weak attractor, as illustrated in Figure 2, where, for the specified initial conditions, the distribution stabilizes at about three-quarters punishers and one-quarter cooperators. The intuition is that with enough punishers around, free riders are very likely to be punished and they, therefore, do worse than cooperators and the many punishers who interact with other punishers or cooperators. Furthermore, it can be shown that elements of this components are "drift compatible" in the sense of Binmore and Samuelson (1999).[4]

So far our model provides evolutionary microfoundations for punishing behavior within a group. The first two theories of punishment in social dilemma experiments were developed to explain the reasons for this punishing behavior. We describe these theories in detail in the next section, but before that discussion we will develop our model further to illustrate foundations for the third theory of punishment.

2.2 Microfoundations for generalized punishment

Now imagine that nature randomly selects four agents at a time instead of two to populate two two-person groups playing the game in Figure 1 in parallel. The two public goods are symmetric, but players derive benefits only from the public good provided within their own two-person group. So far, the structure provides the same equilibria as the baseline game. However, now we let players punish free riders in both their own *ingroup* and in the other *outgroup.* The question we are interested in is whether punishment strategies that are generalized (i.e., not group-specific) can survive or proliferate within a social dilemma institutional structure. More specifically, we wonder whether a simpler heuristic of contribute to your own public good and punish *all* free riders you see, regardless of group affiliation, is viable under very restrictive conditions in which players receive no benefit from punishing in another group, nor do they get higher payoffs in some boundedly rational sense because they conserve the cognitive cost of identifying ingroup and outgroup members.

After changing the structure of the game, we add two more behavioral types to account for the fact that players who cooperate can now punish free riders in both groups. We call the first new type *Pure Social Reciprocators* because they contribute and punish outgroup free riders only. Similarly, *Social Reciprocators* contribute and punish free riders in both groups. We also suppose, for the sake of convenience, that contributors who punish cannot "pick and choose" which means that a contributor who punishes both in- and outgroup players and is matched with three free riders (one ingroup and two outgroup), for example, is assumed to punish all three.

[4] The reader should not be worried that these theoretical results are in some sense "rigged" by restricting the set of strategies as we have done. Our results are consistent with those found in the models discussed in the papers listed in footnote 3 which allow various other strategies. Again, our current purpose is to motivate the evolutionarily viability of punishing strategies not to provide a canonical proof of their existence.

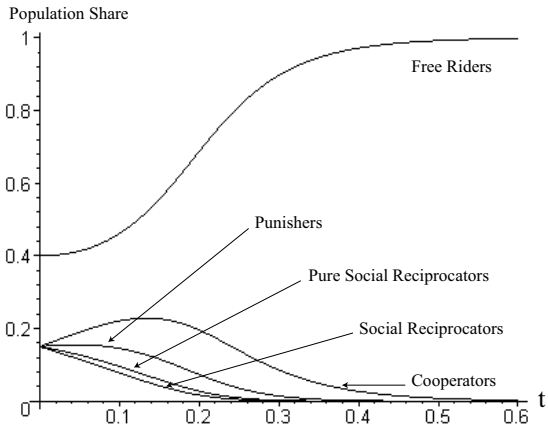

Fig. 3. The evolution of punishing behavior from a balanced population

Like our baseline game, the *mini social reciprocity game* has two Nash equilibria, but a unique subgame perfect equilibrium, in which no one contributes and no one punishes. The second Nash equilibrium is another component in which all players randomize over our four contribution strategies, such that the likelihood of punishment of free riding exceeds some lower bound. For details see Carpenter and Matthews (2002). The expected payoffs are slightly more complicated in the mini social reciprocity game because one needs to account for the fact that free riders can now be punished by outgroup members and that punishers can spend 10, 20 or 30 EMUs on punishment. However, the payoffs to the other contributing strategies are as straight forward as in the baseline game because intuition suggests that the expected payoffs of the contributing behavioral types in the mini social reciprocity game should be a function of the proportion, p_{FR}, of free riders alone, and the numbers bear this out:

$$\pi_C = 75 - 37.5p_{FR}$$
$$\pi_P = 75 - 47.5p_{FR}$$
$$\pi_{PSR} = 75 - 57.5p_{FR}$$
$$\pi_{SR} = 75 - 62.5p_{FR}$$

It can also be shown that the payoff to free riding is:

$$\pi_{FR} = 27.5 + 22.5p_{FR} + 60p_C + 40p_P + 20p_{PSR}$$

Because free riders will sometimes do much worse than punishers, socially reciprocal strategies may evolve from certain initial conditions. Figure 3 illustrates the evolution of behavioral types from the initial state in which 40% of the population are free riders and the remaining strategies each comprise 15% of the population. In this case there are too many free riders for punishment to take hold and despite the cooperators initially doing well by free riding on the punishers, free riders eventually take over the population.

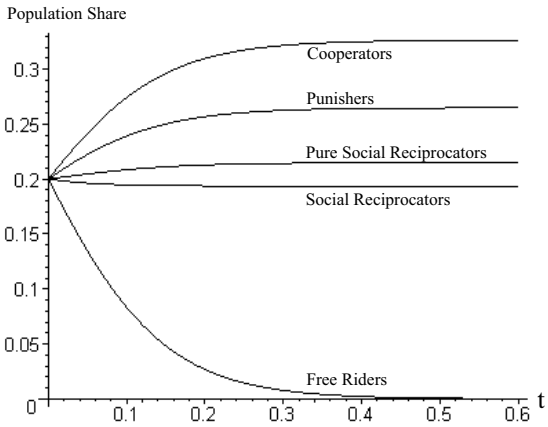

Fig. 4. The dynamics of the social reciprocity game with too many free riders

In a balanced initial population, however, the four contributing types expect $\pi_C = 67.5$, $\pi_P = 65.5$, $\pi_{PSR} = 63.5$, and $\pi_{SR} = 62.5$, and free riders expect just $\pi_{FR} = 56$. The mean payoff for the entire population is 63, which indicates that free riders and social reciprocators will fall short of the population mean, and see their numbers diminish, but the shares of the three other types of contributors will rise. The surprise, perhaps, is that from this initial state, simulation of the replicator dynamic reveals that the free riders and *not* the social reciprocators, will be driven to extinction. The evolution of social reciprocity from balanced initial conditions is shown in Figure 4.[5]

A second important fact is illustrated in Figure 4. Not only do social reciprocators survive selection in a balanced population, so do punishers and cooperators. Perhaps one of the most interesting results of this model is that it predicts a polymorphism of contributing strategies and this polymorphism is, more or less, what we see in the lab. In an experiment that we discuss in Section 4.5, about a third of the participants consistently punish both outside and inside their groups, about half punish ingroup only, and the remaining 20% effectively never punish at all.

3 Why punish?

Section 2 established that punishing strategies may survive in a population, but the models say nothing about why people punish. There are three theories that have been developed to provide the motivation for (or the psychology of) punishment. The common starting point for each theory is the fact that punishing behaviors are grounded in the evolutionary logic provided in Section 2. With respect to our model, the first two theories are represented by the behavioral type we call punishers. The third theory can explain punishers and the two socially reciprocal behavioral types.

[5] As in the simple model, different initial conditions will be associated with different elements of the component and some of these are drift-compatible. However, in general, cooperative outcomes can be sustained as long as $p_P + 2p_{PSR} + 3p_{SR} > 0.625$.

However, the theories differ on two dimensions: the degree to which punishment is believed to be a purposeful act versus a normative response and the discretion that is attributed to punishers in terms of who they punish.

3.1 Fitness Differential Theory (Price et al., 2002)

Price et al. (2002), hereafter PCT, have developed the *fitness differential theory* based on the seemingly obvious assertion (as PCT point out) that no pro-social behavior can evolve in a population of free riders unless cooperative types somehow recover or eliminate the benefits of free riding.[6] Therefore punishment, according to PCT, arises to tax away the benefits accruing to free riders. Specifically, PCT assert that punishment evolves as a *punitive sentiment* which is a desire that the target of the sentiment be harmed. This sentiment is hypothesized to be "hard-wired" into our motivational circuitry. This hard-wired sentiment causes game players to punish free riders even when it is material costly to do so.

PCT also point out that such a punitive sentiment could evolve in response to two possible problems faced by humans in the late Pleistocene: to increase contributions to collective actions and to eliminate the fitness differential that free riders enjoy over cooperators. Their data (discussed below) is consistent only with the hypothesis that game players punish to reduce the payoff advantage accruing to free riders.

PCT are good about proposing how to test whether the fitness differential theory works. According to them (pp. 210–211), the trigger for punitive sentiments is the fact that free riders do better materially than contributors. Observationally, this means that a person's contribution to a public good should correlate with their punitive sentiments because only those who contribute will potentially receive lower payoffs than free riders. Furthermore, the more that one contributes, the more likely one should be to punish and punishment should be directed exclusively from cooperators to free riders. They also state that the relationship between willingness to punish and the willingness to contribute should be robust to the inclusion of other factors that might elicit punishment like the magnitude of the potential benefits from the public good being successfully provided.

At the same time, PCT argue that there should be no relationship between a willingness to reward contributions and one's own contribution decision because rewarding under-providers restores the payoff differential that is suppose to trigger punitive sentiments. Although ingroup and outgroup distinctions are not addressed in PCT, it is clear that payoff-concerned punishers would always choose ingroup punishment over outgroup punishment because spending resources to punish outgroup can only reduce one's relative position within a group.

[6] It is not actually true that free rider benefits need to be taxed away for cooperation to evolve. For example, assortative interactions allow for the evolution of cooperation by restricting access to the gains from cooperation.

3.2 Strong reciprocity (Gintis, 2000)

Reciprocity-based punishment behaviors are similar in phenotype to the behavior posited by PCT (free riders are punished and their fitness is reduced), but the reasons for acting are different. People motivated to reduce fitness differentials are outcome-oriented while, we believe, reciprocators respond to the actions or intentions of others.

One reciprocity-based theory, originating in Gintis (2000), and discussed in the context of economic experiments in Gintis et al. (2003) shows that a behavior called *strong reciprocity*, which is a predisposition to cooperate and punish free riding at some personal cost within well-defined groups, even if the groups have uncertain futures.[7] According to the strong reciprocity theory, the trigger for punishment is the fact that free riders display ill intentions by violating the group-beneficial contribution norm not the fact that they accrue higher payoffs.

Because strong reciprocators care about norm violations and the intentions of the people they play with more than payoffs, they should react differently to the costs of punishment than people who care mostly about payoff differentials. For example, although strong reciprocators may be sensitive to the cost of punishing free riders, they should not care whether the punishment of a free rider causes her payoff to fall below the cooperator payoff. In contrast, the fitness differential theory would suggest that players will only punish when doing so reduces the payoff of the free rider more than her own payoff.

Like the fitness differential theory, strong reciprocity is not equipped to explain the sort of cross-group dynamics we see in Section 2.2. Punishing outgroup doesn't make sense when norms are defined over within-group behavior.

3.3 Social reciprocity (Carpenter and Matthews, 2002)

A second reciprocity-based theory used to explain punishment is called social reciprocity. This theory generalizes the notion of strong reciprocity to account for cooperation and punishment in the sort of large, amorphous, groups that constitute neighborhoods, for example.

We define *social reciprocity* as the act of demonstrating one's disapproval, at some personal cost, for the violation of a widely-held norm, regardless of the material consequences. Social reciprocators, like strong reciprocators, are motivated more by intentions than by payoff differentials, but social reciprocators differ from strong reciprocators because they punish all norm-violators regardless of group affiliation (recall our results in Sect. 2.2).

[7] The model in Gintis (2000) is based on group selection wherein selection happens at both the group level and the individual level so that prosocial traits can evolve if they cause groups to expand and splinter off at rates greater than the internal decay caused by free riding. However, our theoretical results in Section 2 indicate that group selection is not required for the evolution of punishment strategies that are observationally equivalent to strong reciprocity. This fact is also motivated in Bowles and Gintis (2003) using a different mechanism under which free riders are punished by being ostracized from the group.

Table 2. A taxonomy of punishment in social dilemmas

	Punish ingroup only	Punish ingroup and outgroup
Punish for payoff reasons	Fitness differential theory	–
Punish for normative reasons	Strong reciprocity	Social reciprocity

In everyday life, social reciprocity explains the sort of altruistic acts that social psychologists termed the *responsive bystander phenomenon* after the research of Bibb Latane and John Darley. In one of their classic experiments, Latane and Darley (1970), subjects were brought into a room under the pretence of waiting for an interview to begin and paid their show-up fees. However, also in the room was a confederate of the researchers who stole some of the left over show-up money when the experimenter left the room. In 50% of the cases in which subjects report having seen the theft, they reported it to the experimenter. Notice that this is a costly act because the thief might retaliate and there is no expectation that the reporter will receive any benefit from doing so. The idea of social reciprocity is that people will intervene at some cost to themselves when people break obvious rules, and they do so regardless of which group the norm-violator belongs to or whether or not doing so will benefit the social reciprocator in any way.

The description of social reciprocity indicates that if punishment is driven by this motivation, people will punish in two scenarios that differentiate them from fitness differential punishers and strong reciprocators. First, social reciprocators (like strong reciprocators) will punish even when doing so will not remove the payoff differential that free riders achieve. Second, unlike strong reciprocators, social reciprocators will punish norm violations that occur outside their groups.

3.4 Differentiating theories of punishment

The differences in the three theories of punishment can be summarized using Table 2. There are two dimensions on which the theories differ. Two theories, strong reciprocity and social reciprocity, posit intentions that are based in the violation of prosocial norms while the fitness differential theory explains intentions in terms of wanting to eliminate the payoff differential achieved by free riders. Hence, using data that sheds light on the motivations of punishers we can discriminate between the reciprocal theories and the fitness differential theory.

On the other dimension, only social reciprocity explains the punishment that happens outside of one's immediate group. Occurrences of out-group punishment are situations in which punishers or their groups can not receive any future benefits and free riders impose no costs directly on them. In these situations fitness differential players would not punish because doing so is costly and would reduce their relative payoff within their own group. The cost imposed on fitness differential players who punish outgroup reduces their relative payoff compared to both cooperators and free riders within their group and if free riders in the other group contribute more in the future, players in the other group will do better compared to

the altruistic punisher. In sum, fitness differential punishers would always do better by substituting ingroup punishment for outgroup punishment. Likewise, strong reciprocators would not punish outside their groups because their behavior is defined to be regulated by ingroup norms only.

In the next section we examine the behavioral relevance of each theory based on the data from two vignette studies and three laboratory experiments. We examine the vignette study done by the authors of the fitness differential theory (Price et al., 2002) and a more detailed vignette study we conducted specifically to differentiate normative concerns from payoff concerns. As for lab data, we discuss two treatments from Carpenter and Matthews (2002) which link punishment data with contribution behavior and we examine the effect of changing the ratio of punishment costs to the target's payoff reduction based on the data collected in Carpenter (2002a).

4 Survey and experimental evidence of punishing behaviors

To test which theory of punishment is the most consistent with the existing data, we discuss the data from five surveys and experiments. We do not attempt a comprehensive review of the punishment literature; instead our purpose is to discuss only those studies that tell us something about player motivations to punish free riders (in terms of outcomes versus norm violations) and those studies that test whether or not punishment generalizes beyond obvious group boundaries.

4.1 The survey of Price, Cosmides, and Tooby (2002)

PCT presented vignettes about the US going to war and needing to draft soldiers to 287 undergraduates to read and then asked them their opinions about four statements that had to do with their willingness to participate in the war, their willingness to punish free riders, their personal interest in the collective goal of winning the war, and their willingness to reward participants. Both vignettes described warfare between the US and another country but, to balance the design, half the participants read an offensive vignette and the other half read a defensive vignette. The vignettes and statements appear in the Appendix.

According to the fitness differential theory, there should be a link between a respondent's willingness to serve and her willingness to punish free riders because the more likely one is to serve, the larger one's expected payoff will drop below a free rider (see Sect. 3.1). Indeed, PCT find uncontrolled correlations of 0.60 ($p < 0.001$) and 0.65 ($p < 0.001$) and partial correlations of 0.55 ($p < 0.001$) and 0.62 ($p < 0.001$) when controlling for the respondent's reported personal interest in the success of the war in the defensive and offensive scenarios, respectively. This establishes a link between one's willingness to contribute and one's willingness to punish free riders.

PCT also posit that one's willingness to participate should not predict one's willingness to reward other contributors because rewarding other people can only increase fitness differentials. This hypothesis fails the test of simple correlations ($\rho = 0.12$ and $p < 0.05$ for both scenarios), but is resurrected when the authors

control for the respondent's perceived interest in the war. In the latter case, neither partial correlation is larger than 0.03 nor are they significantly different than zero.

The fact that participation predicts punishment and does not predict rewards in this survey study is the primary evidence offered by PCT to support their fitness differential theory.

4.2 Mutual monitoring experiments

In the typical public goods experiment (e.g., Isaac et al., 1984), participants are randomly assigned to groups of four people, given an endowment of money, and given the opportunity to contribute as much of their endowments as they want to a public good. The experiment usually lasts between ten and twenty rounds and the participants are either shuffled into new groups at the beginning of each round (the *strangers* treatment) or stay in the same groups for the entire experiment (the *partners* treatment). The incentives are the same as those discussed in Section 2 and do not depend on the grouping mechanism: each player has a dominant strategy to free ride on the contributions of the other players. Specifically, the payoff function is:

$$\pi_i = (w - c_i) + m\left(\sum c_i\right)$$

where w is each players endowment, c_i is the amount contributed by player i, and $0< m <1$ is the marginal per capita return on the public good. Differentiating the individual return by c_i indicates that free riding dominates, and differentiating the sum of the individual payoffs by the sum of the c_is indicates that contributing is socially efficient – hence the game is a social dilemma.

In a *mutual monitoring* experiment, a second stage is added to the game in which players are shown the (anonymous) contributions of the other group members and given the opportunity to punish them at some cost. One such experiment is reported in Carpenter and Matthews (2002) in which 96 undergraduates participated (24 in the mutual monitoring treatment) and earned $16.55, on average. In this experiment players had a per-period endowment of 25 EMUs, contributions returned 0.5 EMUs for each group member, and punishers paid 1 EMU to destroy 2 EMUs of their target's gross income for the period. Therefore, the payoff function becomes:

$$\pi_i = (25 - c_i) + 0.5\left(\sum c_i\right) - \sum p_{ij} - 2\sum p_{ji}$$

where p_{ij} is the punishment that player i inflicts on player j.

The average contributions in this game are presented in Figure 5 alongside the contributions in a control treatment in which punishment was not allowed. As one can see, punishment tends to stabilize or increase initial contribution levels while contributions decline over time in the standard voluntary contribution game.

We can try to verify the PCT survey results by looking at the relationship between punishing and contributing in this experiment. In the equations reported below the number of EMUs spent on punishment is regressed on the target's level of free riding which is the number of EMUs the target kept normalized by the

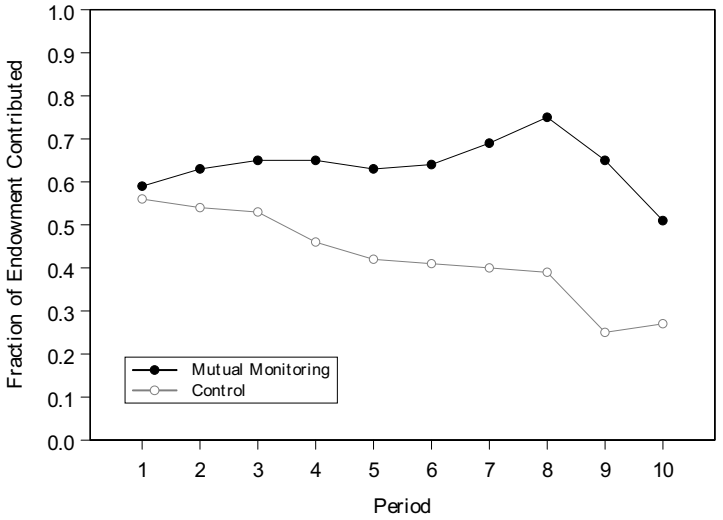

Fig. 5. The evolution of social reciprocity in a balanced population

endowment and the punisher's level of free riding. Specifically, $Punishment_{i,j}$ is the number of EMUs that punisher i spent to destroy j's money, $FreeRide_j$ is a measure between zero and one of how much j free rode, and $FreeRide_i$ is how much the punisher, i, free rode. According to the fitness differential theory, the coefficient on $FreeRide_i$ should be negative and significant because PCT hypothesize that only cooperators punish free riders. The first equation is simple ordinary least squares.[8] Here we see a strong relationship between free riding and being punished (i.e., the coefficient on $FreeRide_j$ is highly significant, $p < 0.001$), but no significant relationship between contributing one's self and punishing others ($p = 0.41$), although the sign of the coefficient is correct.

$$\begin{aligned} Punishment_{i,j} &= 0.14 + 2.54 FreeRide_j - 0.25 FreeRide_i + \varepsilon \qquad (1)\\ &\qquad (0.15)\ \ (0.30) \qquad\qquad (0.31)\\ n &= 720, R^2 = 0.10, F = 40.38\\ Punishment_{i,j} &= -7.44 + 9.16 FreeRide_j - 1.94 FreeRide_i + \varepsilon \qquad (2)\\ &\qquad (0.93)\ \ (0.98) \qquad\qquad (1.14)\\ n &= 720, Wald\chi^2 = 87.67 \end{aligned}$$

However, controlling for individual heterogeneity over time with random effects and accounting for the fact that punishment can not be less than zero using the Tobit estimator in regression 2, the coefficient on the punisher's FreeRide regressor increases substantially. Despite this increase in the coefficient, the effect is still only marginally significant ($p = 0.09$) indicating a weak relationship, at best between cooperating one's self and punishing free riders. In sum, we find only weak support for the fitness differential theory in laboratory data based on the test suggested by PCT.

[8] Standard errors of the estimates are presented in parentheses below the coefficients.

However, perhaps PCT do not offer the best test of fitness differential theory.[9] If people are primarily motivated by the differential payoff accruing to free riders, then what should be important is not necessarily the level of free riding by punisher but the difference in free riding between the punisher and the target. Specifically, if the target free rides more than the monitor, the monitor should punish to remove the payoff differential. However, if the monitor free rides more, there should be no reason to punish because the monitor is already doing better, materially. To examine this looser interpretation of fitness differential theory we can regress the punishment a target receives on the difference between the FreeRide of the monitor and the target. Further, we split this difference into two regressors: *MonitorFRLess* takes the absolute value of the difference in FreeRide when the punisher free rides less than the target and *MonitorFRMore* takes the absolute value of the difference when the punisher free rides more than the target. If fitness differential theory is supported, we expect the coefficient on MonitorFRLess to be positive and significant and the coefficient on MonitorFRMore to be statistically indistinguishable from zero. Regression 3, which also includes random effects and uses the Tobit estimator, reports our results.

$$\begin{aligned} Punishment_{i,j} = & -5.90 + 13.92 MonitorFRLess + 2.47 MonitorFRMore 0 + \varepsilon \\ & (0.84)\ (1.35) \qquad\qquad (1.48) \\ & n = 720, Wald\chi^2 = 111.81 \end{aligned} \tag{3}$$

Under this looser interpretation of PCT, the fitness differential theory performs better. The coefficient on *FRLess* is positive and highly significant ($p < 0.01$) indicating players punish the other players who free ride more, a fact that is consistent with taxing away the differential payoff accruing to people who free ride more than the monitor does. However, we also see suggestive evidence that players are willing to punish those other players who free ride less than they do because the *FRMore* coefficient is positive and marginally significant ($p = 0.10$). This, of course does not jibe with the fitness differential theory.

4.3 A team production survey

We conducted another vignette study to test the bounds of the fitness differential theory that was framed in a way that we thought would be more realistic for participants. In our survey, participants responded to statements about a vignette that described a team production scenario in which someone free rides. The vignette and the relevant statements appear below.

> You and a number of other newly hired people are employed by an auto manufacturer and assigned to work in teams of four. Everyone on the team is paid equally and the pay level is determined entirely by how many cars your work team produces. On the first day of work, you and the other three members of your team divide up the production tasks equally. Over the course of the next month, you and two other members of your group work regularly and hard. However, the fourth member of

[9] One of our referees suggested the following.

> the team often hides in a storage room and reads a book instead of working on cars. This means the other three of you must work harder to make the same number of cars as the other four-person teams. At the end of the month, you and everyone else in your group earn the same amount of money.

Our 70 participants were undergraduates at Middlebury College. Each participant was randomly given one version of our survey. In total there were six versions of the survey that balanced the sequence in which the scenario and the questions were presented to remove any order effects. We asked the participants a few demographic questions, had them read the team production vignette in which one member of a team free rides, and then had them respond to statements about the vignette. Two statements are import for our current purpose. The appear below.

> Punish: *Although it might be costly, I would confront the team member who did not work.*
>
> Norm versus Payoff: *I would be more likely to confront the non-working team member because he broke an unstated rule than because I will reduce the benefits he gets from not working.*

The first we will call *Punish*, asked the participants if they would intervene, even if it were costly to do so, to stop a person from free riding. The second statement called *Norm versus Payoff* asked respondents the degree to which they agreed with a statement that said they were more likely to punish free riders because of normative reasons than because of payoff reasons.[10]

Ninety percent of our respondents said they would confront the free rider to some extent. In reality, this number would probably be smaller, but if the hypothetical nature of this question emboldens everyone to the same degree, we can still learn something about people's motivations from this data because we are mostly interested in explaining the variation in responses. Because our questions were marked on a likert scale we use ordered logit regressions to analyze the variance in people's stated willingness to punish free riders. The results are summarized in regression 4:

$$\begin{aligned} Punish = {} & 0.30NvP + 0.23Age - 0.51Female - 0.43EcClasses + \varepsilon \quad (4) \\ & \quad (0.14) \qquad (0.25) \qquad (0.53) \qquad\quad (0.43) \\ n = 70, & \; pseudoR^2 = 0.11, \chi^2 = 13.69 \end{aligned}$$

NvP is the extent to which the respondent said they were more influenced by norm breaking than by reducing the free riders payoff, *EcClasses* is the number of economics classes the respondent had taken, and the other two controls are self-explanatory.

As one can see, not much difference in responses is explained by the controls. Older participants are more likely to punish, women are less likely to punish, and

[10] We also balanced this statement by switching its order in half the cases so that it read, "I would be more likely to confront the team member who did not work because I will reduce the benefits he gets from not working than because he broke an unstated rule" instead of "I would be more likely to confront the non-working team member because he broke an unstated rule than because I will reduce the benefits he gets from not working."

studying economics is associated with reducing one's prosocial tendencies (as was also demonstrated in Carter and Irons, 1991), but none of these effects is significant. However, people who say they are motivated more by norms than by the payoff differential accruing to free riders are more likely to punish ($p = 0.03$). This evidence favors the normative theories of punishment over the payoff differential theory.

4.4 The demand for punishment

In Carpenter (2002a) and Anderson and Putterman (2003) the authors look at the effect on punishing behavior of changing the ratio of how much punishment costs to how much it harms the target. These experiments are particularly useful because they include treatments in which the price of punishment is very high and therefore the cost to punish is larger than the harm inflicted on free riders. These treatments allow us to differentiate among the two reasons for punishment using experimental data. Clearly, players who are motivated to reduce fitness or payoff differentials would never punish when the cost to harm ratio is greater than one. Therefore, the fitness differential theory predicts that the demand for punishment will fall to zero when the cost to harm ratio rises above one. However, while strong and social reciprocators may be price sensitive, they will still punish despite the high price and should therefore have smooth demand functions.

Because the two experiments provide very similar results, we consider only the data from Carpenter (2002a) which we have access to. In this modification of the mutual monitoring experiment described in Section 4.2, the price of destroying one EMU of another player varies according to the following two sequences: {4,2,1,0.5,0.25} or {0.25,0.5,1,2,4}. The first sequence represents the decreasing price treatment and the second sequence represents the increasing price treatment. There were 72 participants who earned approximately $26 each. Half the participants played the decreasing price game and the other half played the increasing price game to balance any order effects. The experiment was 15 periods long which means that the participants were exposed to each price for three periods. From a design point of view, what is important here is that participants were asked to choose punishment levels when the price was 2 and 4. The first case implies that the monitor would have to pay 2 EMUs to destroy 1 EMU of the target and the second case implies that the monitor would have to pay 4 EMUS to destroy one EMU. Clearly, punishing under these price levels does not reduce the payoff differential between the monitor and the target, it increases it.

To examine whether fitness differential theory is supported by the data from this experiment (i.e., punishment falls two zero when the cost is greater than the harm imposed), Figure 6 presents averages of the ratio of the punishment purchased over the number of EMUs kept by the target of punishment for each price level. We divide punishment by the amount kept to control (removing the few cases which involve division by zero) for the target's level of free riding. As one can see, the average amount of punishment does mostly decrease as the cost to harm ratio increases, but it does not fall to zero at the price of one where one can not remove a fitness differential by punishing. In fact, in each price treatment the average punishment per

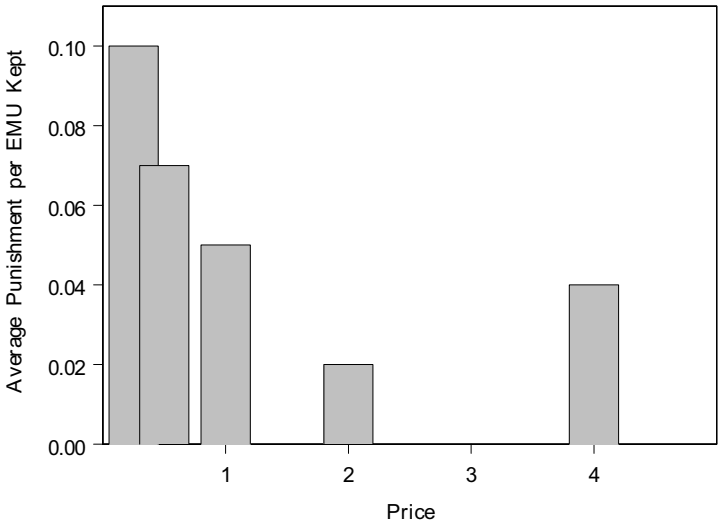

Fig. 6. Average contribution rates in the mutual monitoring public goods game (Source: Carpenter & Matthews, 2002)

EMU kept is significantly greater than zero at better than the 0.01 level indicating that punishment does not fall to zero at any price.

The demand for punishment data does not support the fitness differential theory because people appear willing to punish even when it costs them more to do so than the amount of harm they inflict on the target. This fact, however, is consistent with the reciprocity-based theories of punishment. Calculating the price elasticity of demand for punishment at the regressor means yields a value of -0.64 indicating that, while the demand for punishment is downward sloping, it is relatively inelastic. This fact supports the normative theories of punishment because it suggests that people punish without much regard to the price of doing so.

4.5 A social reciprocity experiment

The social reciprocity experiment is based on the mini social reciprocity game described in Section 2.2. It differs from the standard mutual monitoring game because players can monitor and punish players in completely separate groups, as well as, players in their own groups. In this experiment, discussed in detail in Carpenter and Matthews (2002), two groups of four participants play a voluntary contribution game in parallel. Group members only benefit from contributions to their own, separate, public goods, but at the end of the contribution stage each player is shown the contribution decisions of all eight participants in the session and can destroy the earnings of any person that they want to at a cost of one-half EMU per EMU destroyed. At the end of each period, players are shown how much they earned from their own public good, reminded how much they spent on punishment, and shown how much they themselves were punished. However, if they were punished, they did not know whether it came from within their group or from the other group.

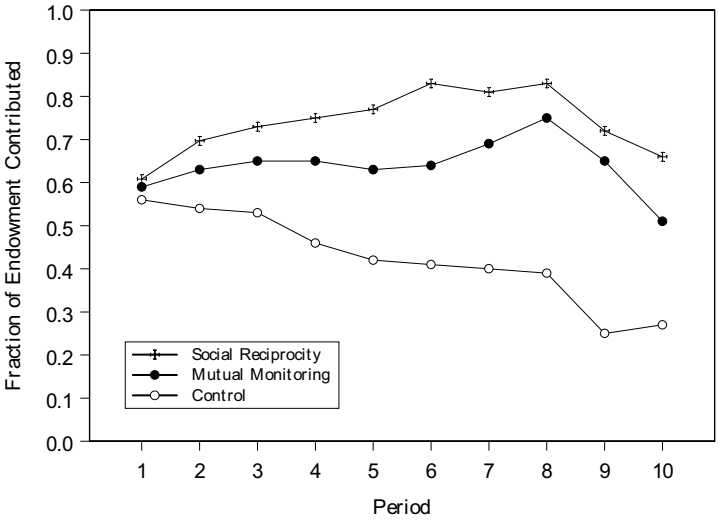

Fig. 7. The effect of changing the price of punishment (Source: Carpenter, 2002)

This feature was added to remove any retaliatory or strategic reasons for future punishment.

As mentioned is Section 2.2, there is a unique subgame perfect equilibrium in the mini social reciprocity experiment that does not include any punishment. We already know that players have no incentive to punish within their groups. Because there is no possibility for a return from punishment in the other "outgroup," egoistic preferences would never have the incentive to do so. Again, because punishing is strictly dominated, free riders need not fear it and, therefore, we should expect complete free riding and no punishment in- or outgroup.

Remember that fitness differential punishers or strong reciprocators will never punish outside their groups which means that any outgroup punishment can only be explained by social reciprocity. Fifty-six Middlebury College undergrads participated in the social reciprocity treatment in 14 groups and 7 sessions. The social reciprocity contribution levels compared to the mutual monitoring game and the no-punishment control are presented in Figure 7. As one can see, the social reciprocity game elicits the highest contributions and, pooling across periods, the contribution levels are statistically different in each game.[11]

There is a good reason why contributions are higher in the social reciprocity experiment – there is more punishment per act of free riding. Fifty percent of our social reciprocity participants punished outside their group at least once. Figure 8 summarizes the expenditures levels of people who punish free riders. It appears that there is more money spent on punishment in the mutual monitoring treatment than on ingroup punishment in the social reciprocity game, but remember that contributions were lower in that game and, therefore, punishment was needed more. The second thing to notice about punishment expenditures is that people in the social

[11] This assertion is based on pairwise means tests and cumulative distribution tests. The smallest t-value was 3.95 and the lowest Kolmogorov-Smirnov statistic was 0.16 ($p < 0.01$).

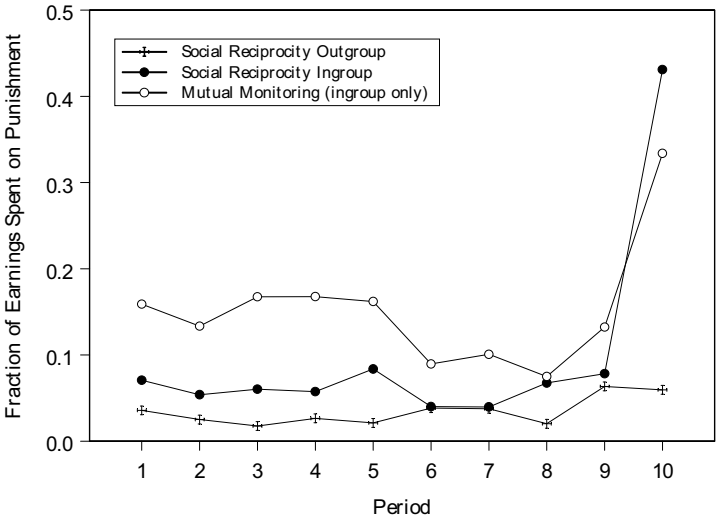

Fig. 8. Contributions in the social reciprocity experiment (Source: Carpenter & Matthews, 2002)

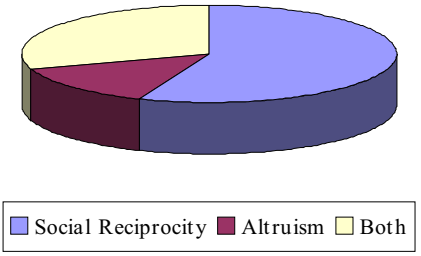

Fig. 9. Punishment levels in the social reciprocity experiment compared to the mutual monitoring experiment (Source: Carpenter & Matthews, 2002)

reciprocity game spend more money punishing ingroup than outgroup.[12] However, the most important result, for our current purposes, is that outgroup punishment is significantly different from zero.[13] Only social reciprocity can explain the behavior of half the participants who punish outgroup and we know this behavior is not error-driven because expenditures on outgroup punishment are positive in each period.[14]

At the end of the social reciprocity experiment we asked people why they punished outside their groups. Figure 9 graphs the distribution of responses. Respondents could choose between altruistic reasons (i.e., punish to increase the contributions in the other group), reciprocal reasons (i.e., punish to get back at rule breakers), or both reasons. Only 14 percent punish for altruistic reasons. The majority of people (56%) punish purely for normative reasons, and 86% of punishment is motivated to some degree by normative concerns.

[12] Pooling the data across periods yields $t = 2.15, p = 0.03$ and $ks = 0.03, p = 0.03$.

[13] $t = 8.57, p < 0.01$.

[14] This sort of outgroup punishment appears to be very robust. It has recently been replicated in a distribution game. See Fehr and Fischbacher (2003).

5 Reconciling the punishment data with theory

We have summarized and provided microfoundations for three theories of punishment in public goods experiments. The three theories differ on at least two dimensions: the extent to which punishment is motivated by payoff concerns versus normative concerns and whether or not the theory can explain punishment that crosses clear group boundaries. The fitness differential theory of Price et al. (2002) posits people punish to remove the payoff differential accruing to free riders and therefore people motivated by these differentials should punish free riders in their groups only when doing so removes the added payoff they receive. Strong reciprocity (Gintis (2000), asserts that people punish free riders without much regard to the cost of punishment to enforce pro-social cooperation norms. This means the strong reciprocators punish rule breakers, but only within their groups. Social reciprocity, developed first in .Carpenter and Matthews (2002), is a generalization of strong reciprocity that says the people subscribe to the simpler heuristic of punishing all norm violators, regardless of group affiliation.

To assess which theory describes behavior best, we have presented the data from five studies. The fitness differential theory is consistent with the fewest of these sources of data. Strong reciprocity performs better than the fitness differential theory, but only social reciprocity is consistent with all the data presented. Price et al. note that fitness differential punishers will be cooperators who direct punishment at free riders because cooperators receive lower payoffs that free riders. However, if reducing payoff differentials is the motivation for punishment, fitness-minded punishers will only punish when the cost of doing so is lower than the harm inflicted on the target. We have found that in at least two sets of data punishment is doled out primarily by cooperators, however, we also see that the price elasticity of punishment is relatively low indicating that people do not based their punishment decisions primarily on the ratio of cost to harm inflicted. In addition, when put "head to head" in a survey similar to the original Price et al. survey, normative reasons for punishment explain much more of the variance in behavior that do payoff reasons. Overall, the only support for the fitness differential theory is the fact that most punishment comes from cooperators and it is directed at free riders.

Both strong reciprocity and social reciprocity are consistent with punishment coming from cooperators and directed at free riders because free riders are seen as norm violators and contributors are norm enforcers. Strong reciprocity and social reciprocity are also consistent with the fact that people punish free riders even when the costs exceed the harm done to the target. However, only social reciprocity can explain the data described in Section 4.5. In this experiment players have the opportunity to monitor and sanction people in different, distinct, groups and about half of the players do so. Hence, social reciprocity is the most parsimonious and robust theory of punishment.

Appendix – The Price, Cosmides, and Tooby (2002) vignettes and statements

Defensive Scenario: Imagine that a few years from now, the Russian people elect a new, warlike dictator who claims that Alaska should rightfully belong to Russia. Under this dictator, Russia invades and conquers Alaska. There is good evidence that Russia also intends to conquer more US territory, in addition to Alaska. In response to the invasion, the USA declares war on Russia. But because the war was unexpected, the USA has allowed its army to get relatively small, and it must start drafting US citizens in order to have a chance of winning the war. How would you feel about the war?

Offensive Scenario: Imagine that several years from now, several oil-rich Middle Eastern countries get together and decide that to increase profits, they will dramatically raise the price of their oil. This price increase devastates US industry and causes high inflation in the USA. US gas prices triple, and several US oil companies go bankrupt. After talks with these Middle Eastern countries fail, the USA declares war on them. But the war was unexpected, so the USA has allowed its army to get relatively small, and it must start drafting US citizens in order to have a chance of victory. How would you feel about the war?

- *If the USA won this war, it would be very good for me as an individual.*
- *If I got drafted, I would probably agree to serve.*
- *If a US citizen resisted this draft, I think they should be punished.*
- *If a drafted US citizen agreed to serve in the war, I'd think they should be rewarded.*

References

Acheson J (1988) The lobster gangs of Maine. University Press of New England, Hanover

Anderson C, Putterman L (2003) Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Brown University Department of Economics Working Paper 2003-15

Axelrod R (1984) An evolutionary approach to norms. American Political Science Review 80: 1095–1111

Binmore K, Samuelson L (1994) An economist's perspective on the evolution of norms. Journal of Institutional and Theoretical Economics 150: 45–63

Binmore K, Samuelson L (1999) Evolutionary drift and equilibrium selection. Review of Economic Studies 66: 363–393

Bochet O, Page T, Putterman L (2003) Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior & Organization (forthcoming)

Bowles S, Carpenter J, Gintis H (2001) Mutual monitoring in teams: theory and evidence on the importance of residual claimancy and reciprocity. Mimeo

Bowles S, Gintis H (2003) The evolution of strong reciprocity. Theoretical Population Biology (forthcoming)

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

Carpenter J (2002a) The demand for punishment. Middlebury College Department of Economics Working Paper #43

Carpenter J (2002b) Punishing free-riders: How group size affects mutual monitoring and collective action. Mimeo

Carpenter J, Matthews P (2002) Social reciprocity. Middlebury College Department of Economics Working Paper

Carter J, Irons M (1991) Are economists different, and if so, why? Journal of Economic Perspectives 5: 171–177

Fehr E, Fischbacher U (2003) Third party punishment and social norms. Institute for Empirical Research in Economics Working Paper, University of Zurich

Fehr E, Gaechter S (2000) Cooperation and punishment in public goods experiments. American Economic Review 90: 980–994

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

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

Gueth W, Kliemt H (1993) Competition or cooperation: On the evolutionary economics of trust, exploitation, and moral attitudes. Metroeconomica 45: 155–187

Isaac R M, Walker J, Thomas S (1984) Divergent evidence on free-riding: An experimental examination of possible explanations. Public Choice 43: 113–149

Latane B, Darley J (1970) The unresponsive bystander: Why doesn't he help? Appleton-Century-Crofts, New York

Masclet D, Noussair C, Tucker S, Villeval M-C (2003) Monetary and nonmonetary punishment in the voluntary contributions mechanism. American Economic Review 93: 366–380

Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge

Ostrom E, Gardner R, Walker J (1994) Rules, games and common-pool resources. University of Michigan Press, Ann Arbor

Price M E, Cosmides L, Tooby J (2002) Punitive sentiment as an anti-free rider psychological device. Evolution and Human Behavior 23: 203–231

Sefton M, Shupp R, Walker J (2000) The effect of rewards and sanctions in provision of public goods. Mimeo

Sethi R (1996) Evolutionary stability and social norms. Journal of Economic Behavior and Organization 29: 113–140

Taylor P, Jonker L (1978) Evolutionary stable strategies and game dynamics. Mathematical Biosciences 40: 145–156