

# Strong Reciprocity and Team Production: Theory and Evidence\*

Jeffery Carpenter,<sup>†</sup> Samuel Bowles<sup>‡</sup>  
Herbert Gintis,<sup>§</sup> and Sung-Ha Hwang<sup>¶</sup>

March 2, 2009

## Abstract

Punishment of shirkers is often an effective means of attenuating incentive problems and sustaining coordination in work teams. Explanations of the motivation to punish generally rely either on small group size or on a Folk theorem that requires coordinated punishment, and hence highly accurate information concerning the behavior of each player. We provide a model of team production in which the punishment of shirkers depends on *strong reciprocity*: the willingness of some team members to contribute altruistically to a joint project and also to bear costs in order to discipline fellow members who do not contribute. This alternative does not require small group size, complex coordinated punishing activities, or implausible informational assumptions. An experimental public goods game provides evidence for the behavioral relevance of strong reciprocity, and how it differs from unconditional altruism.

*JEL classification:* C72; C92; H41

*Keywords:* public good, experiment, punishment, strong reciprocity, team production

---

\*We would like to thank Mark Howard for assistance with the experiment, two referees for thoughtful comments and the European Science, Russell Sage, and National Science Foundations, along with the Behavioral Sciences Program at the Santa Fe Institute and the University of Siena, for research support.

<sup>†</sup>Middlebury College and IZA.

<sup>‡</sup>Santa Fe Institute and University of Siena.

<sup>§</sup>Central European University and Santa Fe Institute.

<sup>¶</sup>University of Massachusetts, Amherst.

## 1 Introduction

The punishment of shirkers by peers in work teams, credit associations, partnerships and local commons situations is often an effective means of attenuating incentive problems that arise where individual actions affecting the well being of others are not subject to enforceable contracts. Explanations of the incentives to punish (Varian, 1990; Stiglitz, 1993) generally rely either on the small size of the interacting group, or on repeated interactions and low discount rates, motivating the application of a Folk theorem (Fudenberg et al., 1994). Neither of these is completely satisfactory, since work teams are not always small and the Folk theorem depends on repeated interactions, coordinated punishment, and hence highly accurate information concerning the behavior of each player (Gintis, 2009) Ch. 10. Other treatments do not address the incentive to confer benefits on other team members by contributing to production and punishing shirkers (Arnott, 1991; Weissing and Ostrom, 1991).<sup>1</sup>

We provide a model of team production in which the motivation to punish is *strong reciprocity*: the willingness of some altruistic team members to engage in the costly punishment of shirkers. The key conditions supporting altruistic punishment are (a) contributing to production becomes a norm, and (b) there are members who punish violators of norms even when this is costly. We call this *altruistic punishment* (Fehr and Gächter, 2002) and those practicing it, who also contribute to production, *strong reciprocators* (Gintis, 2000; Boyd et al., 2003). We call the incentive structure in which cooperation in a team is maintained through altruistic punishing by strong reciprocator members, *self-policing*.

An experimental public goods game provides evidence for the behavioral relevance of strong reciprocity in teams. In a treatment approximating a one-shot interaction, and even on the terminal round of the game, shirkers are punished. We also find that shirkers respond to punishment by increasing their level of cooperation. This experiment replicates the results of Fehr and Gächter (2000) in that there is a positive level of punishment in all periods, and the level of cooperation does not decay when costly punishment is allowed. We also test new hypothe-

---

<sup>1</sup>Dong and Dow (1993a) and Legros and Newman (1996) assume the team can impose collective sanctions on shirkers. This assumption is reasonable if shirking is easily detected and team members have more effective or lower cost forms of punishment than are available to a traditional firm. We do not make this assumption. Dong and Dow (1993b) and Dong and Dow (1993a) assume shirking can be controlled by the threat of non-shirkers to exit the team. However the threat of exiting is credible only if team members have very high fallback positions—in Dong and Dow’s model, this takes the form of independent production—which generally is not the case.

ses derived from the comparative statics of our model of strong reciprocity and team production. With respect to sanctions, we show that the level of punishment directed towards team members increases with the rate at which they shirk, that punishment is increasing in the cost imposed on the team by the shirker, but that punishment decreases as team size grows. In addition, we estimate novel measures of altruism and reciprocity and show that altruists tend to punish less and reciprocators punish more. Considering contributions, we provide evidence that contributions increase with the productivity of the public good but may decrease with increasing team size.

## **2 Strong Reciprocity and Altruistic Punishment in Teams**

The problem of free riding in teams has been addressed by two standard models. The first, due to Alchian and Demsetz (1972), holds that residual claimancy should be assigned to an individual designated to monitor team members' inputs, thus ensuring the incentive compatibility for the (non-contractible) activity of policing itself, while addressing the members' incentive to free ride by the threat of dismissal by the residual claimant monitor. They contrast this view of the 'classical firm,' as they call it, with an alternative in which team members are residual claimants and policing is performed, if at all, by salaried personnel. Alchian and Demsetz correctly observe that group residual claimancy would dilute incentives, but simply posit the allocational superiority of the classical firm: "we assume that if profit sharing had to be relied on for all team members, losses from the resulting increase in central monitor shirking would exceed the output gains from the increased incentives of other team members not to shirk." (1972:786) As we will see, their invocation of the so-called "1/n problem" to justify this assumption is not entirely adequate.

The second approach, pioneered by Holmström (1982), demonstrates that in principal multi-agent models one can achieve efficiency or near-efficiency through contracts that make individual team members residual claimants on the effects of their actions without conferring ownership rights on them. Contracts of this type typically impose large penalties for shirking and require large lump-sum up-front payments on the part of agents, or they pay each team member the entire team output minus a large constant and thus, in the presence of stochastic influences on output, entail negative payments in some periods, or at best a substantial variance of income for team members. These arrangements are infeasible if team members have insufficient wealth. Moreover, where contributions (e.g., work effort) are

continuously variable these incentive mechanisms support large numbers of Nash equilibria, thus rendering breakdown of cooperation likely.

Alchian and Demsetz (1972), Holmström (1982) and related models explain why self-policing in teams is unnecessary. But, the limited applicability of the owner-monitor and optimal contracting approaches provides one motivation for exploring the relationship between residual claimancy and self-policing in teams. Another motivation is empirical. There is some evidence that group residual claimancy is effective, by comparison with payments unrelated to group output, even in quite large teams (Ghemawat, 1995; Hansen, 1997; Knez and Simester, 2001). Self-policing based on residual claimancy appears to be effective in the regulation of common pool resources such as fisheries, irrigation, and grazing lands (Ostrom, 1990), in the regulation of work effort in producer cooperatives (Greenberg, 1986; Craig and Pencavel, 1995) and in the enforcement of non-collateralized credit contracts (Banerjee et al., 1994). Experimental studies (Frohlich et al., 1998) provide additional support for the effects of residual claimancy in inducing lower supervision costs and higher productivity in (small) work teams. Further, the fact that residual claimancy may provide incentives for self-policing even in quite complex settings and large groups is suggested by evidence that in the United States home ownership is a significant predictor of participation in community organizations (Glaeser and DiPasquale, 1999) and local politics but, significantly, not national politics (Verba et al., 1995), as well as willingness to sanction coresidents who transgress social norms (Sampson et al., 1997).

Making team members residual claimants can have positive incentive effects, since team members may have privileged access to information concerning the activities of other team members, and may have the means of disciplining shirkers and rewarding hard work that are not available to third parties. As residual claimants, moreover, team members may have the incentive to use this information and exercise their sanctioning power, even if the team is large. Thus while Alchian and Demsetz are surely correct in saying that residual claimancy in large teams does not substantially reduce the direct incentive to free ride, it may support superior means of sanctioning and hence discouraging free riding through altruistic punishing. Our experiments suggest the motive to punish team members includes a positive utility attached to punishing wrong-doers, independently of any expectation of material gain which might accrue to the punisher as a result of modification of the subsequent behavior of the punished shirkers. Punishing is costly, however, and if the desire to punish is not sufficiently widespread, we shall see, self-policing will fail.

The willingness to engage in costly punishment provides a basis for linking

residual claimancy with altruistic punishment, even in large teams. An individual who shirks inflicts harm on the other members of the team if (and only if) they are residual claimants. Members may then see this violation of reciprocity as reason to punish the shirker. Our model requires only that a considerable fraction of team members be reciprocators. This is in line with the evidence from experimental economics, which indicates that in virtually every experimental setting a certain fraction of the subjects do not act reciprocally, either because they are self-interested, or they are purely altruistic.<sup>2</sup>

A key element in our approach, one shared by recent contributions of Kandell and Lazear (1992), Rotemberg (1994), Banerjee et al. (1994), and Besley and Coate (1995) is that our model is based on ‘social preferences’ which, while unconventional, are well supported by recent experimental and other research.

We assume that though team members observe one another in their productive activity, they cannot design enforceable contracts on actions because this information is not verifiable (cannot be used in courts). In this situation we show that the assignment of residual claimancy to team members can attenuate incentive problems, because it gives strong reciprocators an incentive to punish shirking team members even in large teams.

Two common characteristics of successful self-policing are uncontroversial: the superior information concerning non-verifiable actions of team members available to other team members and the role of residual claimancy in motivating members to acquire and use this information in ways that enhance productivity. Less clear is whether residual claimancy motivates costly punishing in large groups.<sup>3</sup>

An answer consistent with the suite of behaviors we call strong reciprocity is suggested by findings in  $n$ -player public goods experiments. These provide a motivational foundation for self-policing in teams whose members are residual claimants, since these experiments show that agents are willing to incur a cost to punish those whom they perceive to have treated them or members of a group to which they belong badly.<sup>4</sup> In these experiments, which allow subjects to punish non-cooperators at a cost to themselves, the moderate levels of contribution typi-

---

<sup>2</sup>For an especially clear example, see Fischbacher et al. (2001) or Burks et al. (forthcoming). Fehr and Schmidt (1999) provides a survey.

<sup>3</sup>The problem of motivating the punishment of shirkers would not arise, of course, if team members were sufficiently altruistic towards teammates. In this case members would simply internalize the benefits conferred on others by their monitoring. Rotemberg (1994) develops a model of this type. However were team members sufficiently altruistic to motivate mutual monitoring in this way, there would be no initial free rider problem either.

<sup>4</sup>See Ostrom et al. (1992) on common pool resources, Fehr et al. (1997) on public goods.

cally observed in early play often rise in subsequent rounds to near the maximal level, rather than declining to insubstantial levels as in the case where no punishment is permitted. It is also significant that in the experiments of Fehr and Gächter (2000), punishment levels are undiminished in the final rounds, suggesting that disciplining norm violators is an end in itself (deQuervain et al., 2004) and hence will be exhibited even when there is no prospect of modifying the subsequent behavior of the shirker or potential future shirkers (Walker and Halloran, 2004; Carpenter and Matthews, 2005).

Consider a team of size  $n > 2$ , where member  $i$  supplies an amount of effort  $1 - \sigma_i \in [0, 1]$ . We call  $\sigma_i$  the *level of shirking* of member  $i$ , and write  $\bar{\sigma} = \sum_{j=1}^n \sigma_j / n$  for the average level of shirking. We assume shirking at level  $\sigma_i$  adds  $q(1 - \sigma_i)$  dollars to team output, where  $q > 1$ , while the cost of working is a quadratic function  $(1 - \sigma)^2/2$ . We call  $q$  the *productivity of cooperation*. We assume the members of the team share their output equally, so member  $i$ 's payoff is given by

$$\pi_i = q(1 - \bar{\sigma}) - (1 - \sigma_i)^2/2. \quad (1)$$

The payoff loss to each member of the team from one member shirking is  $\beta = q/n$ . We assume  $1/n < \beta < 1$ .

We assume member  $i$  can impose a cost on  $j \neq i$  with monetary equivalent  $s_{ij}$  at cost  $c_i(s_{ij})$  to himself. The cost  $s_{ij}$  results from public criticism, shunning, ostracism, physical violence, exclusion from desirable side-deals, or another form of harm. We assume that acts of punishment, like work effort, are non-verifiable and hence not subject to contract. We also assume  $c_i(0) = c_i'(0) = c_i''(0)$  and  $c_i(s_{ij})$  is increasing, strictly convex, and with  $c_i'''(s_{ij}) \geq 0$  for all  $i, j$  when  $s_{ij} > 0$ .

Member  $j$ 's standing as a cooperative member of the group,  $b_j$ , depends on  $j$ 's level of shirking and the harm that  $j$  inflicts on the group, which we assume is public knowledge. Specifically, we assume

$$b_j = \beta(1 - 2\sigma_j) \quad (2)$$

so  $b_j = -\beta$  if  $j$  completely shirks, and  $b_j = \beta$  if  $j$  does not shirk at all. This means that  $\sigma_j = 1/2$  is the point at which  $i$  evaluates  $j$ 's cooperative behavior as neither good nor bad. This point could be shifted to another value between 0 and 1 by treating  $b_j$  as a quadratic in  $\sigma_j$ , but the added generality is not illuminating.

To model cooperative behavior with both altruistic and reciprocal preferences, we say that individual  $i$ 's utility depends on his own material payoff  $\pi_i$  and the

payoff  $\pi_j$  to other individuals  $j \neq i$  according to:

$$u_i = \pi_i + \sum_{j \neq i} [(a_i + \lambda_i b_j)(\pi_j - s_{ij}) - c_i(s_{ij})] - s_i(\sigma_i) \quad (3)$$

where  $b_j$  is  $j$ 's standing as a cooperative member of the group (which may be either positive or negative),  $s_i(\sigma_i) = \sum_{j \neq i} s_{ji}(\sigma_i)$  is the punishment inflicted upon  $i$  by other group members, and  $\lambda_i \geq 0$ . The parameter  $a_i$  is  $i$ 's level of unconditional altruism if  $a_i > 0$  and unconditional spite if  $a_i < 0$ , and  $\lambda_i$  is  $i$ 's strength of reciprocity motive, valuing  $j$ 's payoffs more highly if  $j$  conforms to  $i$ 's concept of good behavior, and conversely (Rabin, 1993; Levine, 1998). If  $\lambda_i$  and  $a_i$  are both positive, the individual is termed a strong reciprocator, motivated to behave generously towards individuals about whom he knows nothing, but willing to reduce the payoffs of an individual who reveals a bad type even at a cost to himself when his reciprocal preferences outweigh his unconditional altruism.

We assume that individuals maximize (3). Because  $b_j$  can be negative, utility maximization may lead  $i$  to reduce his own contribution, or to punish  $j$ . Note that this motivation for punishing a shirker values the punishment *per se* rather than the benefits likely to accrue to the punisher if the shirker responds positively to the punishment. Moreover, members derive utility from punishing the shirker, not simply from observing that the shirker was punished. This means that punishing is *warm glow* rather than instrumental towards affecting  $j$ 's behavior.<sup>5</sup>

In support of this analytical model, we report below an experiment carried out by the authors involving a public goods game with costly punishment. This experiment replicates the results of Fehr and Gächter (2000), in that there is a positive level of punishment in all periods, and the level of cooperation does not decay when costly punishment is allowed. We also test new hypotheses derived from the comparative statics of our model of strong reciprocity and team production. With respect to sanctions, we show that the level of punishment directed towards team members increases with the rate at which they shirk, that punishment is increasing in the cost imposed on the team by the shirker, but that punishment decreases as team size grows. In addition, we demonstrate that novel measures of altruism and reciprocity correlate to some degree with punishment: Altruists tend to punish less and reciprocators punish more. Considering contributions, we confirm levels of unconditional altruism and conditional cooperation and provide evidence that

---

<sup>5</sup>Of course these other motives for punishment may be important, but the evidence for the existence of 'warm glow' punishment is convincing (Casari and Luini, 2007).

suggests that contributions increase with the productivity of the public good but may decrease in team size.

Member  $i$  will choose  $s_{ij}^*(\sigma_j)$  to maximize utility in (3), giving rise to the first order condition (assuming an interior solution)

$$c_i'(s_{ij}^*) = \lambda_i \beta (2\sigma_j - 1) - a_i, \quad (4)$$

or the marginal cost of punishing is equal to the marginal benefit of reducing  $j$ 's payoffs given  $i$ 's assessment of  $j$ 's type. When  $\lambda_i = 0$  and  $a_i < 0$ ,  $i$  punishes  $j$ , but independent of  $j$ 's shirking level. If  $\lambda_i > 0$  and

$$\sigma_j \leq \sigma_i^0 = \frac{1}{2} \left[ \frac{a_i}{\lambda_i \beta} + 1 \right], \quad (5)$$

the maximization problem has a corner solution in which  $i$  does not punish. We assume  $(a_i/\lambda_i\beta + 1)/2 > 0$ . For  $\lambda_i > 0$  and  $\sigma_j > \sigma_i^0$ , denoting the right hand side of (4) by  $\phi$ , and differentiating (4) totally with respect to any parameter  $x$ , we get

$$\frac{ds_{ij}^*}{dx} = \frac{\partial \phi}{\partial x} \frac{1}{c_i''(s_{ij}^*)}. \quad (6)$$

In particular, setting  $x = a_i$ ,  $x = \lambda_i$ ,  $x = \sigma_j$ ,  $x = \beta$  and  $x = n$  in turn in (6), we see that

*Theorem 1. For  $\lambda_i > 0$  and  $\sigma_j > \max\{1/2, \sigma_i^0\}$ , the level of punishment by  $i$  imposed on  $j$ ,  $s_{ij}^*$ , is (a) decreasing in  $i$ 's unconditional altruism  $a_i$ ; (b) increasing in  $i$ 's reciprocity motive,  $\lambda_i$ ; (c) increasing in the level  $\sigma_j$  of  $j$ 's shirking; (d) increasing in the harm  $\beta$  that  $j$  inflicts  $i$  by shirking; and (e) decreasing in group size.*

The punishment  $s_j(\sigma_j)$  inflicted upon  $j$  by the group is given by

$$s_j(\sigma_j) = \sum_{i \neq j} s_{ij}^*(\sigma_j), \quad (7)$$

which is then nondecreasing in  $\sigma_j$ , and is differentiable where it is strictly positive. From (5) and the assumption  $(a_i/\lambda_i\beta + 1)/2 > 0$ , we see that  $s_i(0) = s_i'(0) = 0$  for all  $i$ .

The first order condition on  $\sigma_i$  from (3) is given by

$$1 - \sigma_i - \beta - \beta \sum_{j \neq i} (a_i + \lambda_i b_j) - s_i'(\sigma_i) = 0, \quad (8)$$

so  $i$  shirks up to such point as the net benefits of shirking equal  $i$ 's valuation of the cost imposed on others by his shirking plus the marginal cost of shirking entailed by the increased level of punishment that  $i$  may expect. This defines  $i$ 's optimal shirking level  $\sigma_i^*$  for all  $i$ , and hence closes the model. We assume  $s_i''(\sigma_i) > -1$ , so the second order condition for a maximum is satisfied.

We say that  $i$ 's partners *shirk on balance* if  $\sigma_j > 1/2$  for all  $j \neq i$  and they *work on balance* if the opposite inequalities hold. We then have the following theorem, whose proof is in the Appendix.

*Theorem 2. Suppose there is an stable interior equilibrium under a best response dynamic. Then (a) an increase in  $i$ 's unconditional altruism  $a_i$  leads  $i$  to shirk less, and (b) an increase in  $i$ 's reciprocity motive  $\lambda_i$  leads  $i$  to work more when  $i$ 's partners work on balance.*

### 3 Experimental Evidence

Our model embodies an essential behavioral assumption, namely that under some conditions strong reciprocity will induce sufficient punishment of free riding to sustain high levels of team output. To test the plausibility of this assumption, we conducted an experimental public goods game, extending the standard protocol by making each player's contribution to the public good known to all team members at the end of each round, and allowing players to punish other players based on this information, at a cost to themselves. Fehr and Gächter (2000) used a similar experimental setting to show that there is indeed a propensity to punish, and that allowing costly punishment in a multiperiod setting prevented the decay of cooperation usually found in public goods experiments (see Ledyard (1995)).<sup>6</sup> In addition to replicating Fehr and Gächter, we investigate the comparative statics developed in Theorems 1 and 2 above.<sup>7</sup>

We deliberately created an experimental environment in which contributions would be difficult to sustain by implementing the so-called *strangers treatment*, in which subjects are randomly reassigned to a new group at the beginning of each round of play.<sup>8</sup> We also make punishing shirkers costly; the cost of inflicting a

---

<sup>6</sup>Other contributions to the literature on the ability of sanctions to control free riding in social dilemma settings include Barr (2001), Cinyabuguma et al. (2006), Masclet et al. (2003), and Walker and Halloran (2004).

<sup>7</sup>The instructions for a typical session appear in the Appendix.

<sup>8</sup>The more common *partners treatment*, in which groups remain together throughout the experiment tends to foster more cooperation than the stranger treatment (Croson, 1996; Fehr and

penalty of two experimental monetary units, EMUs, is one EMU for the punisher.

Suppose there are  $n$  players, each player receives an endowment of  $w$  EMUs at the beginning of each round, and player  $i$  contributes  $w(1 - \sigma_i)$  to the public good. These contributions are revealed to the players, who then can punish the others.<sup>9</sup> Let  $s_{ij}$  be the expenditure on sanctions assigned by player  $i$  to player  $j$  (we force  $s_{ii} = 0$ ). Then the payoff to player  $i$  is

$$\pi_i = w [\sigma_i + q(1 - \bar{\sigma})] - \sum_{j=1}^n s_{ij} - 2 \sum_{j=1}^n s_{ji}, \quad (9)$$

where  $\bar{\sigma} = \sum_{j=1}^n \sigma_j/n$  is the average shirking rate,  $\sum_j s_{ij}$  is player  $i$ 's expenditure on sanctions and  $2 \sum_j s_{ji}$  is the reduction in  $i$ 's payoffs due to the total sanctions received from the rest of the team. Note that the familiar MPCR (marginal per capita return on the public good) is just  $q/n$ . Also, the unit endowment in the model developed in the previous two sections is  $w = 25$  EMUs in our experiment.

To examine how subjects' contributions and punishment allocations are affected by team size, the productivity of the public good and the degree of harm caused by shirking we used the design of Carpenter (2007) with two group sizes, four and eight, and two values of  $q/n$ , 0.3 and 0.7. This design allows us to compare across treatments for similarities in behavior based on the cost that shirkers inflict on their teammates.<sup>10</sup> While Carpenter (2007) focused on the effect of the monitoring group size (the fraction of the other team members that each person can monitor and punish) we focus on differences in actual group size and productivity of the public good. In addition, while our design is similar, Carpenter (2007) used the punishment technology pioneered in Fehr and Gächter (2000) in which punishers removed shares of a target's income in 10% increments and we used the more straightforward institution by which punishment expenditures are multiplied by some constant factor to determine the harm inflicted on each target.

A total of twelve sessions were conducted (three per treatment) with 172 participants. Figure 1 summarizes our experimental design. The number of participants, and therefore teams, per treatment vary due to no-shows. All subjects were

---

Gächter, 2000; Keser and van Winden, 2000). Each session had either 12 or 16 participants. This means that the realized probability of being rematched in the same group from one round to the next was higher than we would have liked. However, none of our results depend on the strangers design - we implemented it to provide a more stringent test of the theory.

<sup>9</sup>The instructions to participants refer neutrally to "reductions" with no interpretation supplied.

<sup>10</sup>On their own, the effects of group size (Isaac and Walker, 1988; Isaac et al., 1994) and  $q/n$  (Isaac et al., 1984) have been studied extensively, but not in the context of monitoring and punishment.

	Four-person teams	Eight-person teams
$q/n=0.30$	10 teams 40 subjects	6 teams 48 subjects
$q/n=0.70$	9 teams 36 subjects	6 teams 48 subjects

**Figure 1:** Experimental Treatments

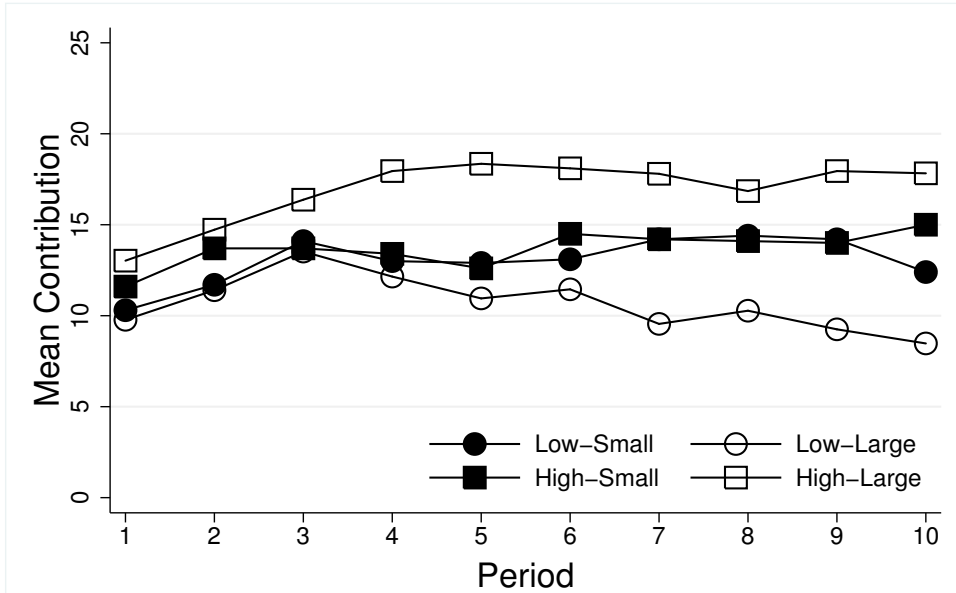
recruited by email from the general student population at the University of Massachusetts (enrollment of approximately 20,000) and none had ever participated in a public goods experiment before. Each subject was given a five dollar show-up fee upon arrival and then was seated at a partially isolated computer terminal so that decisions were made in privacy. Each session took approximately 45 minutes from sign-in to payments and subjects earned \$20.58 on average, including the show-up fee.

Each session lasted ten periods. In each period (a) subjects were randomly reassigned to a group, given an endowment of  $w = 25$  EMUs, and allowed to contribute, anonymously, any fraction of the endowment to a public account, the remainder going to the subject's private account; (b) the total group contribution, the subject's gross earnings, and the contributions of other team members (presented in random order) were revealed to each subject, who was then permitted to assign sanctions to others; (c) payoffs were calculated according to (9), and subjects were informed of their net payoffs for the period.

If we assume standard self-regarding preferences for participants (i.e. each player cares only about his or her personal payoff) then punishing is not a best response because it is costly and any benefits will be diffuse. Hence, the unique subgame perfect Nash equilibrium of the game is that no one punishes and therefore no one contributes to the public good.

The results of our experiments conform to all of the comparative static predictions in Theorems 1 and 2. We proceed with our analysis of the evidence as follows. First, we provide an overview of the experimental data. We then describe how Figure 2 addresses the pattern of general individual decisions. Clearly, the aggregate contributions in the current experiment do not exhibit the decline seen in

<sup>11</sup>In each case we report the results of Tobit regressions (because both punishment and contributions can be censored) that include individual random effects (to control for cross-sectional heterogeneity) and time period fixed effects.



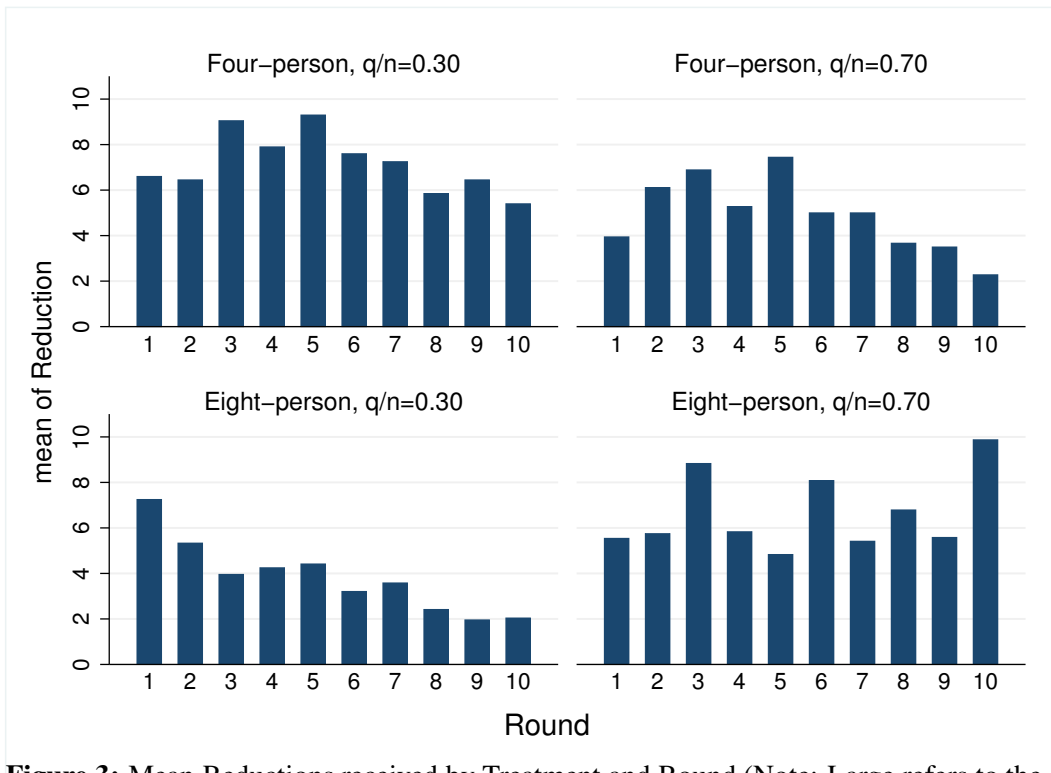
**Figure 2:** Mean Contribution by Treatment and Round (Note: Low or High refers to the team’s  $q/n$  and Small or Large refers to the team size).

the standard public goods game (Ledyard, 1995). In this sense our results look similar to other studies of the public goods game with punishment (e.g., Fehr and Gächter, 2000; Masclet et al., 2003; Carpenter and Matthews, 2005). Average contributions pooled across treatments start at 45% of the endowment in period one, rise slightly, level off, and end at 54% in period ten. The reason behind sustained levels of cooperation in our experiment is punishment. Overall, 89% of the participants punished another participant at least once.<sup>12</sup>

In Figure 3 we illustrate the mean amount of punishment incurred by participants in the four treatments over the ten rounds of the experiment. The figure indicates that punishment is used widely in each treatment and, although the levels fall in two of the treatments over time, punishment is also used extensively in the last periods of the experiment.<sup>13</sup> Figure 4 shows the probability that a player

<sup>12</sup>The frequency of punishers is higher than Carpenter (2007) which found that 68% of people punished. Our high frequency applies to all treatments: 98% punished in the Low  $q/n$ , Small  $n$  treatment; 88% punished in the Low  $q/n$ , Large  $n$  treatment; 81% punished in the High  $q/n$ , Small  $n$  treatment; and 90% punished in the High  $q/n$ , Large  $n$  treatment.

<sup>13</sup>A number of studies have found that punishment is used even in the last round when it can have

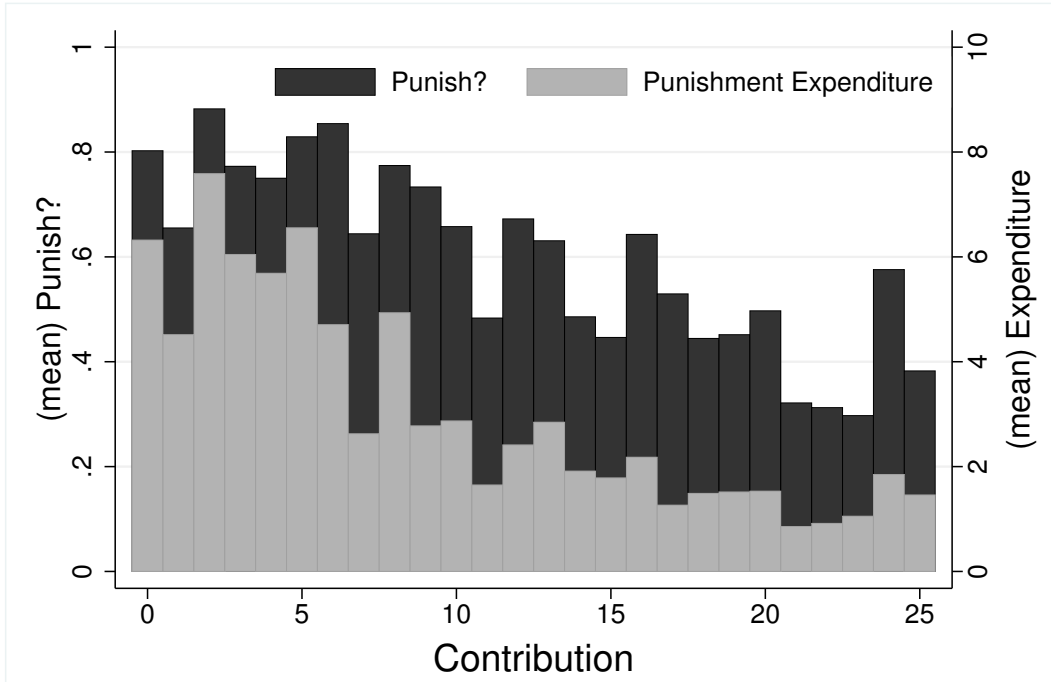


**Figure 3:** Mean Reductions received by Treatment and Round (Note: Large refers to the team size and High refers to the value of  $m$ )

was punished and the mean reduction the player received from one of the other team members based on the target's contribution. The pattern is clear (except for a few anomalies due to small samples; e.g., there were only 17 out of 1720 observations of someone contributing 2 units) those players that shirk more are more likely to be punished and they are punished more severely. In addition, the level of punishment seems to dwindle faster than the likelihood of punishment as targets contribute more.

---

no affect on future behavior (Fehr and Gächter, 2000; Masclet et al., 2003; Walker and Halloran, 2004).



**Figure 4:** Who is Punished? The probability of being punished and the mean reduction subtracted due to a single punisher by contribution choice.

#### 4 Altruism, Reciprocity, Shame, and Spite

Altruism and reciprocity play a central role in the comparative statics derived as Theorems 1 and 2. We use the method developed in Carpenter and Seki (2005) to estimate individual level altruism and reciprocity measures based on the contribution data gathered in our experiment. We then use these measures to evaluate the first parts of our two theorems. However, we begin by describing how we estimate social preference parameters for our participants.

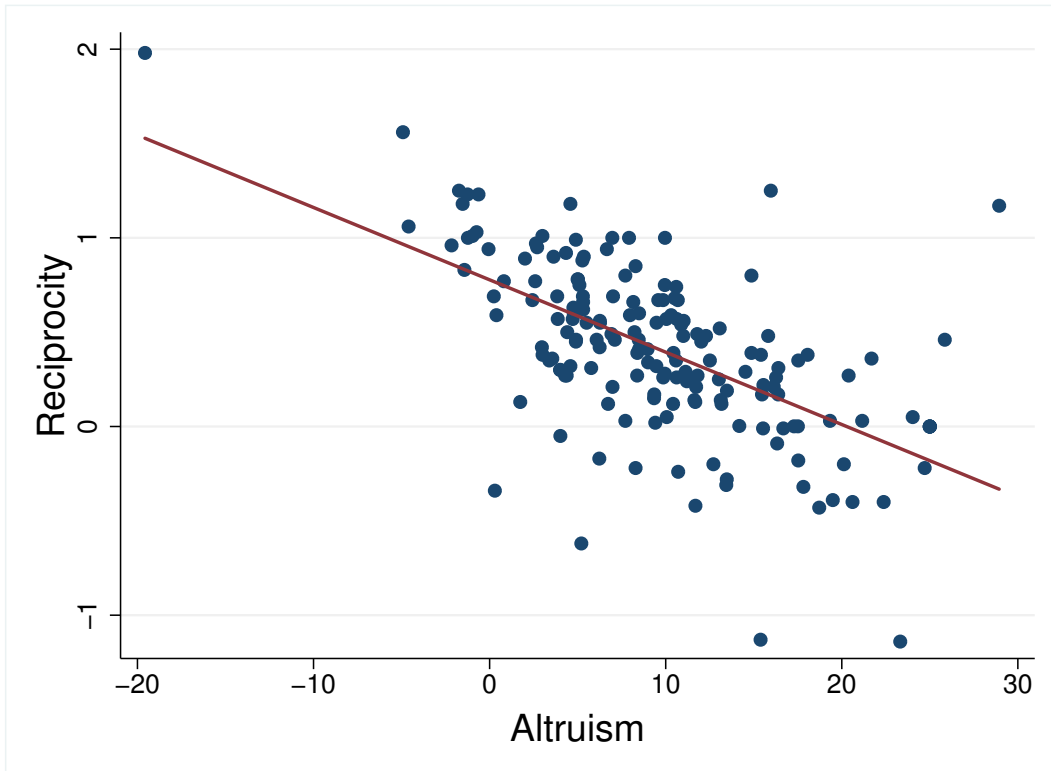
We regress contributions in period  $t$  on the average contribution of the other team members in period  $t - 1$ , the amount of punishment received by the decision-maker in period  $t - 1$  and a constant. At the population level, the coefficient on the average contribution of the others is 0.40 ( $p < 0.01$ ) and the constant is 9.77 ( $p < 0.01$ ) which indicates that, on average, our participants would contribute 9.77 EMUs even if the other group members contributed nothing in the previous period and their contributions are increasing in the contributions of the others. In other words, our average participant appears altruistic because she will contribute a positive amount even if the others contribute nothing and she appears reciprocal

because she will increase or decrease her contribution in response to what the others have done. We therefore interpret the constant in this regression as a measure of unconditional altruism or spite (a linear transformation of  $a_i$  in our model) and the slope as a measure of conditional cooperation or reciprocity (a linear transformation of  $\lambda_i$  in the model). Note that we included the punishment that one received in the regression so that our measures of altruism and reciprocity are not biased by the fact that punishment also affects contributions.

While the overall regression results hint at the average behavioral attributes of our participants, to evaluate the model we need individual measures of altruism and reciprocity. One way of generating these measures is to use the random coefficients estimator (Swamy, 1970), which allows the estimated coefficients to vary by individual. This makes sense because few of our participants demonstrated behavior exactly consistent with the coefficients reported above. Instead some players are altruistic but not very reciprocal, some are reciprocal but not altruistic and some are neither. In the random coefficients model the parameters are assumed to be draws from a population of possible coefficients and these possible coefficients vary from one person to the next. These models are sensible when the data come from the repeated sampling of individual subjects and the model for each subject can be assumed to be a deviation from a broader population model. The output of the model is an intercept and slope for each subject.

The altruism measures vary from -19 to 29 with a mean of 9.59. Although there is significant mass below zero in our reciprocity estimates (indicating that about 20% of the participants react negatively to the contributions of others), the mean estimate of the individual response is 0.41. In addition, 51% of the individual altruism estimates and 41% of the reciprocity measures are significantly different from zero. Figure 5 illustrates the "types" of individuals who participated in the experiment. In general, altruism and reciprocity are negatively correlated ( $\rho = 0.6$ ,  $p < 0.01$ ) but there is considerable variation in the joint distribution of these two characteristics.

Theorem 1 suggests that punishment should: decrease in altruism, increase in reciprocity, increase in the level of the target's shirking, increase in the harm imposed by shirking and decrease in group size. We use the regressions reported in Table 1 to assess these predictions. In column (1) we regress individual punishment choices on our social preference estimates and the other parameters referred to in Theorem 1. All the coefficients are highly significant and carry signs consistent with our theory. To facilitate comparisons, we report standardized regression coefficients for our measures of social preference. A standard deviation increase in altruism and reciprocity reduces the amount spent on punishment by



**Figure 5:** The distribution of observed altruism and reciprocity.

0.13 standard deviations in the first case and increases punishment by 0.31 standard deviations in the second case. For each additional EMU kept by the target, punishment increases by a quarter EMU, as the MPCR increases from 0.3 to 0.7, punishment increase slightly less than one EMU, and for each additional team member, punishment decreases by 0.36 EMUs.<sup>14</sup> Although our measures of altruism and reciprocity should be unbiased, the point estimates are measured within some confidence interval and in columns (2) and (3) of Table 1 we account for the differences in the sharpness of these estimates by weighting our observations by the inverse of the standard errors squared. There are two sets of weights to consider: the altruism standard errors and the reciprocity standard errors. In column (2) we use the altruism weights and in column (3) we use the reciprocity weights. In each case we see that accounting for the size of the confidence interval around our preference measures does not change our estimates by very much and all the coefficients remain significant at conventional levels.

<sup>14</sup>These last two results agree with those described in Table 2 of Carpenter (2007).

	(1)	(2)	(3)
Altruism (estimated)	-0.13 <sup>s</sup> (0.07)**	-0.15 <sup>s</sup> (0.04)***	-0.06 <sup>s</sup> (0.04)*
Reciprocity (estimated)	0.31 <sup>s</sup> (0.06)***	0.47 <sup>s</sup> (0.04)***	0.47 <sup>s</sup> (0.10)***
$\sigma_j w$ (target's level of shirking)	0.25 (0.01)***	0.28 (0.01)***	0.32 (0.01)***
$\frac{q}{n}$ (harm imposed by shirking)	2.24 (0.45)***	1.92 (0.32)***	3.16 (0.33)***
$n$ (team size)	-0.36 (0.05)***	-0.38 (0.03)***	-0.42 (0.03)***
Constant	-3.91 (0.50)***	-4.46 (0.35)***	-5.95 (0.35)***
$\chi^2$	790	2091	2282
$N$	9000	9000	9000

**Table 1:** The Determinants of Punishment. Note: all regressions are Tobits, include random effects and time period fixed effects. The dependent variable is the punishment choice of an individual punisher and standard errors of the estimates are in parentheses. In column (2) observations are weighted by the precision of the altruism measures and in column (3) the observations are weighted by the precision of the reciprocity measures. <sup>s</sup> indicates standardized regression coefficient. \* indicates significant at the 0.10, \*\* 0.05, \*\*\* 0.01 levels.

Concerning shirking, Theorem 2 predicts that: altruism will cause team members to shirk less, and reciprocity will lead members to shirk less if their teammates work on balance. These predictions are examined in Table 2. We begin in column (1) by regressing individual contribution choices on the lagged average contribution of the other team members,  $q$  and  $n$ .<sup>15</sup> It appears that the predictions in Theorem 2 find support in our data. The constant in column (1) is positive, highly significant and evidence of altruistic contributions: with all the other variables "turned off" participants still contribute 6.49 EMUs. We also see that reciprocity motivates contributions. As the average contribution of the other team

<sup>15</sup>Column (1) of Table 2 is similar to the specification used to generate the individual measures of altruism and reciprocity discussed in the context of Table 1. However, the two are not exactly the same because the results reported in Table 2 are more complicated in that they account for the effect of the treatments and employ the Tobit estimator.

members increase by an EMU, the average participant increases her own contribution by almost half an EMU.<sup>16</sup> Lastly, we see that a unit shift in the productivity of the public good increases contributions by 1.43 EMUs, on average, and that contributions decrease by 0.60 EMUs for each additional team member.

In column (2) of Table 2 we extend our contribution results past the implications of Theorem 2 to look at some simple dynamics. We add the lag of one's own contribution and the lag of the punishment that one receives to the specification in column (1). These additions do not change our conclusions about the predictions of Theorem 2 with two exceptions. First, the coefficient on team size shrinks in half and is no longer significant at the 10% level; however the  $p$ -value only rises to 0.13. Second, the constant shrinks to 1.63, which is not too worrisome because it would be odd if the inertia now captured in the lagged dependent variable was not previously captured by the constant. In either case, the results (in this context) are at least symptomatic of altruism. The two additional regressors in column (2) of Table 2 suggest that there is considerable inertia in contributions and that people react pro-socially to punishment. A standard deviation increase in received punishment (4.46) increases one's contribution by 0.58 EMUs, on average.

Does shirking pay once the costs of punishment are considered? Note that the act of shirking deprives the shirker of the returns from the group project, so the net benefit of shirking in the absence of punishment is just  $1 - \frac{q}{n}$  which averages 0.5 EMUs in our experiment. From Table 1 recall that the estimated punishment expenditure is 0.25 for each additional EMU kept which imposes a 0.50 EMU ( $2 \times 0.25$ ) fine on the shirker. In other words, shirkers break even at the margin. What about punishment? Are sanctions a good investment? The punisher invests 0.5 EMUs to remove 1 EMU from the shirker. On average, the shirker will respond by increasing her contribution by only 0.13 EMUs. This does not seem like a lot; however, if this punishment happens at the end of the first period, the 0.13 increase accumulates to 0.21 EMUs by the end of round ten because of the inertial effect captured by the coefficient on the lagged dependent variable.<sup>17</sup> Despite the cumulative effect, punishing does not pay materially (i.e.,  $0.21 < 0.50$ ), although there is mounting evidence that it pays subjectively (de-Quervain et al., 2004; Hopfensitz and Reuben, 2007). Of course the summed benefits to *all* team members more than offset the costs to the punisher (e.g., with

---

<sup>16</sup>Fehr and Fischbacher (2003) find a similar reciprocal relationship between expectations and contributions.

<sup>17</sup>Of course there are other dynamic forces that our analysis does not account for. For example, as one referee pointed out, people might anticipate punishment and contribute more. In most cases, these other forces suggest that our estimate of the effectiveness of punishment is conservative.

teams of four  $0.84 > 0.50$ ). Our results are thus consistent with the interpretation that punishment by strong reciprocators is altruistic in the standard sense that it reduces the actor's payoffs while benefiting others.

To explore the response to punishment, we define a shirker in a given round to be a player who contributed less than the team average in that round. We call "contributors" those players who contribute more than the average (and therefore have negative deviations). We hypothesize that contributors and shirkers respond differently to punishment because the former may feel spite or anger when punished while the latter may feel shame (Bowles and Gintis, 2005). Because the experimental instructions refer to "contributions to a group project," it is soon clear to participants why shirkers are punished, but when it occurs, punishment is probably confusing for contributors.

To study possible differences in responses to punishment, column 3 of Table 2 separates shirkers from contributors and interacts lagged punishment with players' deviations from the average to test whether punishment, *per se*, matters or if punishment only matters when one deviates from the norm. We see that for both free riders and contributors, movement towards the mean is increased by punishment and the effect of punishment is greater the farther one is from the mean. In both cases, free riders move towards the mean even if not punished, but more strongly so when they are. Consistent with the experimental results of Hopfensitz and Reuben (2007), a reasonable interpretation is that social emotions provide the motivation for this behavior. Contributors respond spitefully to being punished, regarding fellow members as being unkind, valuing their payoffs negatively and as a result contributing less. Shirkers experience shame when punished and respond positively.

## 5 Conclusion

Our model and experimental results suggest that under appropriate conditions, strong reciprocity can support the punishment of shirkers and the maintenance of high levels of cooperation, unless the frequency of reciprocators is too low or the group is too large. Furthermore, self-policing supports levels of member effort that are close to first best. The effectiveness of altruistic punishment by peers in raising the contributions of shirkers depends on the fact that the incentives provided by punishment do not crowd out pre-existing social preferences that might have induced contributions in the absence of punishment, as is observed in a large number of public goods and principal agent experiments surveyed in Bowles

(2008) and Bowles and Hwang (2008). The counterproductive effects of explicit incentives in the experiments they survey appear to arise when the punishment or fines fails to evoke shame in the shirker, but rather conveys negative information about the individual imposing the incentive, for example that he is seeking to monopolize the gains from cooperation. The altruistic punishment by peers does not have this crowding out effect because the act of punishment reduces the punisher's payoffs and thereby is more likely to evoke shame in the shirker, with increased subsequent contributions the response.

The case we have modeled is a production team in which the noncontractible action is work effort. But the model may equally depict a range of analogous problems. The 'team' might be composed of family, neighbors and friends engaged in informal insurance to supplement market-supplied insurance as in Stiglitz (1993) and Arnott (1991), or members of an informal borrowing group where team members borrow from a financial institution with a renewal of credit being contingent on all members repaying at the end of the first round. Both cases conform to the assumptions of the above model, namely, superior information held by team members, combined with interdependence of members welfare on other members' actions, and low-cost opportunities to punish members who impose costs on others in the team.

Another application is to residential home owners. Here, the team consists of neighbors whose residential amenities, and hence the value of their housing assets, are affected by the noncontractual actions of others in the neighborhood. Sampson et al. (1997) provide empirical evidence of such self-policing in neighborhoods. In this case, monitoring and punishment may consist of admonitions favoring anything from maintaining the appearance of one's property to joining in collective actions to gain safer streets or better schools for the neighborhood. Finally, we think the model may illuminate a characteristic of the foraging bands which constituted human society during most of its history, namely, widespread hunting, foraging, and food sharing, and punishment of those who violated the underlying reciprocity norms (Knauff, 1991; Boehm, 1993; Wiessner, 2005; Bowles and Gintis, 2004).

Given the apparently widespread nature of the problems of non-contractibility which it addresses and the welfare benefits it may make possible, self-policing ought to be ubiquitous in modern economies. But it is not. A reason suggested by this model is that residual claimancy by team members is essential to the underlying monitoring motivations, and for many of the relevant production teams the fact that members are asset poor effectively precludes assignment of any but trivial levels of residual claimancy to team members. Transferring residual claimancy over

the income streams of an asset but not ownership itself to team members creates incentives for the team to depreciate the assets, the costs of which may more than offset any gains from self-policing. Thus prohibitive costs may arise if residual claimancy is separated from ownership, and outright ownership may be precluded by borrowing limitations and possibly high levels of risk aversion characteristic of low wealth team members.

The role of residual claimancy in motivating self-policing thus provides another case in which differing distributions of wealth may support differing equilibrium distributions of contracts and systems of governance. Other cases include forms of agricultural and residential tenancy (Laffont and Matoussi, 1995) access to self employment and human investment (Loury, 1981; Blanchflower and Oswald, 1998; Black et al., 1996), and the extent of cooperative forms of ownership of team assets (Legros and Newman, 1996).<sup>18</sup> Because efficient contracts and other incentive mechanisms may be implementable under some distributions of wealth but not under others, a particular distribution of wealth may preclude allocationally superior systems of contract and incentives.

## 6 Appendix

We will use the stability condition

$$\sum_{j \neq i} \frac{\partial \sigma_j}{\partial \sigma_i} < 1 \quad (10)$$

with which the model is stable under a best response dynamic, given by

$$\frac{d\sigma_i}{dt} = \sigma_i(\sigma_{-i}) - \sigma_i \quad \text{for } i = 1, \dots, n. \quad (11)$$

with fixed point  $\sigma^* = (\sigma_1^*, \dots, \sigma_n^*)$ . The Jacobian matrix of this dynamical system is  $A - I$ , where  $I$  is the  $n$ -dimensional identity matrix and  $A = (a_{ij})$  where  $a_{ii} = 0$  and  $a_{ij} = \partial \sigma_i / \partial \sigma_j$  for  $i, j = 1, \dots, n, i \neq j$ . If the real parts of the eigenvalues of this matrix are negative, the equilibrium is stable. Our stability assumption assures that  $A^T$  is substochastic, so  $\lim_{n \rightarrow \infty} A^n = 0$ , so every eigenvalue  $\eta$  of  $A$  satisfies  $|\eta| < 1$ . However,  $Av = \eta v$  implies  $(A - I)v = (\eta - 1)v$ , so the eigenvalues of  $A - I$  all have negative real parts, and the dynamical system is stable.

---

<sup>18</sup>We survey these cases in Bardhan et al. (2000).

We now show that assuming stability condition (10), we have

$$\text{sign} \left( \frac{\partial \sigma_1}{\partial x} \right) = \text{sign} \left( \frac{d\sigma_1^*}{dx} \right). \quad (12)$$

for  $x = a_1$  and  $x = \lambda_1$  under the assumed conditions. Differentiating (8) totally with respect to parameter  $x$  and rearranging, we get

$$(I - A) \frac{d\sigma^*}{dx} = \frac{\partial \sigma}{\partial x}. \quad (13)$$

Because the eigenvalues of  $A$  have modulus less than unity,  $I - A$  is invertible and we have  $(I - A)^{-1} = I + A + A^2 + \dots = I + M$ . Because  $A$  is a nonnegative matrix ( $a_{ij} = 2\beta^2\lambda_i > 0$  for  $i \neq j$  and  $a_{ii} = 0$ ), we have  $M > 0$ . Moreover, writing the second order condition corresponding to (4) as  $\psi < 0$ , we have

$$\frac{\partial \sigma_1}{\partial a_1} = \frac{\beta(n-1)}{\psi} < 0$$

and

$$\frac{\partial \sigma_1}{\partial \lambda_1} = -\frac{\beta^2(n-1)}{\psi} (2\bar{\sigma}_{-1}^* - 1),$$

where

$$\bar{\sigma}_{-i}^* = \frac{1}{n-1} \sum_{j \neq i} \sigma_j^*.$$

Now, for  $j > 1$ , totally differentiating (8) with respect to  $a_1$ , we have

$$\frac{\partial \sigma_j}{\partial a_1} = \frac{1}{\psi} \frac{\partial^2 s_j}{\partial \sigma_j \partial a_1}. \quad (14)$$

From (4) and

$$\frac{\partial s_{1j}^*}{\partial a_1} = -\frac{1}{c_1''(s_{1j}^*)},$$

we have

$$\frac{\partial^2 s_{1j}^*}{\partial \sigma_j \partial a_1} = \frac{2\lambda_1 \beta c_1'''}{(c_1'')^3}.$$

Therefore, we have

$$\frac{\partial \sigma_j}{\partial a_1} = \frac{1}{\psi} \frac{2\lambda_1 \beta c_1'''}{(c_1'')^3} \leq 0. \quad (15)$$

We thus have

$$\frac{d\sigma_1}{da_1} = (1 + m_{11})\frac{\partial\sigma_1}{\partial a_1} + m_{12}\frac{\partial\sigma_2}{\partial a_1} + \dots + m_{1n}\frac{\partial\sigma_n}{\partial a_1} < 0,$$

where  $(m_{ij}) = M$ .

Similarly, for  $j > 1$ , totally differentiating (8) with respect to  $\lambda_1$ , we have

$$\frac{\partial\sigma_j}{\partial\lambda_1} = \frac{1}{\psi} \frac{\partial^2 s_j}{\partial\sigma_j \partial\lambda_1}. \quad (16)$$

From (4) and

$$\frac{\partial s_{1j}^*}{\partial\lambda_1} = \frac{\beta}{c_1''(s_{1j}^*)} (2\sigma_j - 1)$$

we have

$$\frac{\partial^2 s_{1j}^*}{\partial\sigma_j \partial\lambda_1} = \frac{2\beta}{c_1''} \left[ 1 - \frac{c_1'''}{(c_1'')^2} \lambda_1 \beta (2\sigma_j^* - 1) \right].$$

Therefore, when 1's partners work on balance, we have

$$\frac{\partial\sigma_j}{\partial\lambda_1} = \frac{1}{\psi} \frac{2\beta}{c_1''} \left[ 1 - \frac{c_1'''}{(c_1'')^2} \lambda_1 \beta (2\sigma_j^* - 1) \right] < 0. \quad (17)$$

Because working on balance implies  $\bar{\sigma}_{-1}^* < 1/2$ , hence  $\partial\sigma_1/\partial\lambda_1 < 0$ , we now have

$$\frac{d\sigma_1}{d\lambda_1} = (1 + m_{11})\frac{\partial\sigma_1}{\partial\lambda_1} + m_{12}\frac{\partial\sigma_2}{\partial\lambda_1} + \dots + m_{1n}\frac{\partial\sigma_n}{\partial\lambda_1},$$

which is negative when player 1's partners work on balance program.

## REFERENCES

- Alchian, A., Demsetz, H., 1972. Production, information costs, and economic organization. *American Economic Review* 62, 777–795.
- Arnott, R., 1991. Moral hazard and nonmarket institutions. *American Economic Review* 81, 180–190.
- Banerjee, A.V., Besley, T., Guinnane, T.W., 1994. Thy neighbor's keeper: The design of a credit cooperative with theory and a test. *Quarterly Journal of Economics*, 491–515.

- Bardhan, P., Bowles, S., Gintis, H., 2000. Wealth inequality, credit constraints, and economic performance. In Atkinson, A., Bourguignon, F. (Eds.), *Handbook of Income Distribution*. North-Holland, Dordrecht, pp. 541–603.
- Barr, A., 2001. Social dilemmas, shame based sanctions, and shamelessness: Experimental results from rural zimbabwe. Working Paper, Oxford University.
- Besley, T., Coate, S., 1995. Group lending, repayment incentives and social collateral. *Journal of Development Economics* 46, 1–18.
- Black, J., de Meza, D., Jeffreys, D., 1996. House prices, the supply of collateral and the enterprise economy. *Economic Journal* 106, 60–75.
- Blanchflower, D., Oswald, A., 1998. What makes a young entrepreneur? *Journal of Labor Economics* 16, 26–60.
- Boehm, C., 1993. Egalitarian behavior and reverse dominance hierarchy. *Current Anthropology* 34, 227–254.
- Bowles, S., 2008. Policies designed for self-interested citizens may undermine “the moral sentiments”: Evidence from economic experiments. *Science* 320.
- Bowles, S., Gintis, H., 2004. The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology* 65, 17–28.
- Bowles, S., Gintis, H., 2005. Prosocial emotions. In Blume, L.E., Durlauf, S.N. (Eds.), *The Economy As an Evolving Complex System III*. Santa Fe Institute, Santa Fe, NM.
- Bowles, S., Hwang, S.H., 2008. Social preferences and public economics: Mechanism design when preferences depend on incentives. *Journal of Public Economics* 92.
- Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. Evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* 100, 3531–3535.
- Burks, S.V., Carpenter, J.P., Goette, L., forthcoming. Performance pay and the erosion of worker cooperation: Field experimental evidence. *Journal of Economic Behavior and Organization* .
- Carpenter, J., 2007. Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior* 60, 31–51.
- Carpenter, J., Matthews, P., 2005. Norm enforcement: Anger, indignation, or reciprocity. Department of Economics, Middlebury College, Working Paper 0503.
- Carpenter, J., Seki, E., 2005. Do social preferences increase productivity? field

- experimental evidence from fishermen in toyama bay. IZA Discussion Paper 1697.
- Casari, M., Luini, L., 2007. Group cooperation under alternative peer punishment technologies: An experiment. Department of Economics, University of Siena.
- Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? *Experimental Economics* 9, 265–279.
- Craig, B., Pencavel, J., 1995. Participation and productivity: A comparison of worker cooperatives and conventional firms in the plywood industry. *Brookings Papers: Microeconomics* , 121–160.
- Croson, R., 1996. Partners and strangers revisited. *Economic Letters* 53, 25–32.
- deQuervain, D.J.F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. *Science* 305, 1254–1258.
- Dong, X., Dow, G., 1993a. Does free exit reduce shirking in production teams? *Journal of Comparative Economics* 17, 472–484.
- Dong, X., Dow, G., 1993b. Monitoring costs in chinese agricultural teams. *Journal of Political Economy* 101, 539–553.
- Fehr, E., Fischbacher, U., 2003. The nature of human altruism—proximate patterns and evolutionary origins. *Nature* 425, 785–791.
- Fehr, E., Gächter, S., 2000. Cooperation and punishment. *American Economic Review* 90, 980–994.
- Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. *Nature* 415, 137–140.
- Fehr, E., Gächter, S., Kirchsteiger, G., 1997. Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica* 65, 833–860.
- Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 817–868.
- Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? *Economic Letters* 71, 397–404.
- Frohlich, N., Godard, J., Oppenheimer, J., Starke, F., 1998. Employee vs. conventionally owned and controlled firms: An experimental analysis. *Managerial and Decision Economics* 19, 311–326.
- Fudenberg, D., Levine, D.K., Maskin, E., 1994. The folk theorem with imperfect public information. *Econometrica* 62, 997–1039.

- Ghemawat, P., 1995. Competitive advantage and internal organization. *Journal of Economic and Management Strategy* 3, 685–717.
- Gintis, H., 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206, 169–179.
- Gintis, H., 2009. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, Princeton, NJ.
- Glaeser, E.L., DiPasquale, D., 1999. Incentives and social capital: Are homeowners better citizens? *Journal of Urban Economics* 45, 354–384.
- Greenberg, E., 1986. *Workplace Democracy: The Political Effects of Participation*. Cornell University Press, Ithica, NY.
- Hansen, D.G., 1997. Individual responses to a group incentive. *Industrial and Labor Relations Review* 51, 37–49.
- Holmström, B., 1982. Moral hazard in teams. *Bell Journal of Economics* 7, 324–340.
- Hopfensitz, A., Reuben, E., 2007. The importance of emotions for the effectiveness of social punishment. Discussion paper 05-075. Tinbergen Institute.
- Isaac, R.M., Walker, J., Thomas, S., 1984. Divergent evidence on free-riding: an experimental examination of possible explanations. *Public Choice* 43, 113–149.
- Isaac, R.M., Walker, J.M., 1988. Group size effects in public goods provision: The voluntary contribution mechanism. *Quarterly Journal of Economics* 103, 179–200.
- Isaac, R.M., Walker, J.M., Williams, A.W., 1994. Group size and voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of Public Economics* 54, 1–36.
- Kandel, E., Lazear, E.P., 1992. Peer pressure and partnerships. *Journal of Political Economy* 100, 801–817.
- Keser, C., van Winden, F., 2000. Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics* 102, 23–39.
- Knauff, B., 1991. Violence and sociality in human evolution. *Current Anthropology* 32, 391–428.
- Knez, M., Simester, D., 2001. Firm-wide incentives and mutual monitoring at continental airlines. *Journal of Labor Economics* 19, 743–772.
- Laffont, J.J., Matoussi, M.S., 1995. Moral hazard, financial constraints, and share cropping in el oulja. *Review of Economic Studies* 62, 381–399.

- Ledyard, J.O., 1995. Public goods: A survey of experimental research. In Kagel, J.H., Roth, A.E. (Eds.), *The Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ, pp. 111–194.
- Legros, P., Newman, A.F., 1996. Wealth effects, distribution, and the theory of organization. *Journal of Economic Theory* .
- Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1, 593–622.
- Loury, G., 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49, 843–67.
- Masclot, D., Noussair, C., Tucker, S., Villeval, M.C., 2003. Monetary and non-monetary 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, UK.
- Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: Self-governance is possible. *American Political Science Review* 86, 404–417.
- Rabin, M., 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83, 1281–1302.
- Rotemberg, J.J., 1994. Human relations in the workplace. *Journal of Political Economy* 102, 684–717.
- Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* 277, 918–924.
- Stiglitz, J., 1993. *Welfare Economics with Asymmetric Information (Lindhal Lectures)*. Oxford University Press, Oxford.
- Swamy, P.A.V.B., 1970. Efficient inference in a random coefficient regression model. *Econometrica*, 38(2), 311–323. 38, 311–323.
- Varian, H.R., 1990. Monitoring agents with other agents. *Journal of Institutional and Theoretical Economics* 46, 153–174.
- Verba, S., Schlozman, K.L., Brady, H., 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press, Cambridge, MA.
- Walker, J., Halloran, M., 2004. Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics* 7, 235–247.
- Weissing, F., Ostrom, E., 1991. Irrigation institutions and the games irrigators play: Rule enforcement without guards. In Selten, R. (Ed.), *Game Equilibrium Models II: Methods, Morals and Markets*. Springer-Verlag, Berlin, pp. 188–262.

Wiessner, P., 2005. Norm enforcement among the ju/'hoansi bushmen: A case of strong reciprocity? *Human Nature* 16, 115–145.

## **7 Appendix - Participant instructions (four person, $m=0.30$ treatment)**

You have been asked to participate in an experiment. For participating today and being on time you have been paid \$5. You may earn an additional amount of money depending on your decisions in the experiment. This money will be paid to you, in cash, at the end of the experiment. By clicking the BEGIN button you will be asked for some personal information. After everyone enters this information we will start the instructions for the experiment.

Please be patient while others finish entering their personal information. The instructions will begin shortly.

During the experiment we will speak in terms of Experimental Monetary Units (EMUs) instead of Dollars. Your payoffs will be calculated in terms of EMUs and then translated at the end of the experiment into dollars at the following rate: 30 EMUs = 1 Dollar.

The experiment is divided into 10 different periods. In each period participants are divided into groups of 4. The composition of the groups will change randomly at the beginning of each period. This means that in each period your group will consist of different participants.

Each period of the experiment has two stages. In the first stage you will decide how many EMUs you want to contribute to a group project. At the second stage of each period you will be shown the contributions of the other members of your group. You will then decide whether you want to reduce the earnings of the other members of your group.

Now we will explain the two stages in more depth.

### **Stage One**

At the beginning of every period each participant receives a 25 EMU endowment. You have to decide how much of the 25 EMUs you want to contribute to the group project and how many you want to keep for yourself. You are asked to contribute whole EMU amounts (i.e. a contribution of 5 EMUs is alright, but 3.85 should be rounded up to 4).

To record your decision, you will type EMUs amounts in two text input boxes, one for the group project labeled **GROUP ALLOCATION** and one for yourself

labeled PRIVATE ALLOCATION. These boxes will be yellow. Once you have made your decision, there will be a green SUBMIT button that will record your decision.

After all the members of your group have made their decisions, each of you will be informed of your gross earnings for the period.

Your Gross Earnings will consist of two parts:

- 1) The EMUs you kept for yourself.
- 2) Your return from the Group Project. Your earnings from the group project equal 0.3 times the total EMUs contributed by all the members of the group.

Your Earnings can be summarized as follows:

$$1 \times (\text{EMUs you keep}) + 0.3 \times (\text{Total EMUs Allocated to the Group Project})$$

Each EMU you keep for yourself benefits you alone. EMUs you allocate to the group project yield a return of 0.3 for you. However, every member of the group receives 0.3 EMUs for each EMU you allocate to the group project. Similarly, you receive a return of 0.3 EMUs for every EMU that other members of the group allocate to the group project. Thus, gross earnings in a period are the number of EMUs you keep for yourself, plus the return from all the EMUs you and other members of the group allocate to the group project.

#### Stage Two

In stage two you will be shown the allocation decisions made by other members of your group and they will see your decision. Also, at this stage you will be able to reduce the earnings made by other members of your group, if you want to, and the other members of your group will be able to reduce your earnings. You will be shown how much each member of your group kept and how much they allocated to the group project. Your allocation decision will also appear on the screen and will be labeled 'YOU'. Please remember that the composition of your group will change at the beginning of each period and therefore it is very unlikely that a person in your group this period will also be in your group next period.

At this point you will decide how much (if at all) you wish to reduce the earnings of the other members of your group. You reduce someone's earnings by typing the number of EMUs you wish to spend to reduce that person's earnings into the input text box that appears below that group member's allocation decision.

You will have to pay a cost to reduce the earnings of other group members. For each EMU you spend you will reduce the earnings of the other group member by 2 EMUs. You can spend as much of your accumulated earnings as you wish to reduce the earnings of each of the other group members.

Consider the case where there are 4 people per group. Suppose you spend 2 EMUs to reduce the earnings of one of the members of your group, you spend

9 EMUs reducing the earnings of another group member, and you don't spend anything to reduce the earnings of the last member of your group. Your total cost of reductions will be  $(2+9+0)$  or 11 EMUs. When you have finished distributing reductions you will click the blue DONE button.

How much a participant's earnings from the first stage are reduced is determined by the total amount spent by all the other members of the group. If a total of 3 EMUs is spent, then his or her earnings would be reduced by 6 EMUs. If the other group members spend 4 EMUs in total, his or her earnings would be reduced by 8 EMUs.

Nobody's earnings will be reduced below zero by the other members of the group. For example, if your gross earnings were 40 EMUs and the other group members spent 22 EMUs to reduce your earnings, your gross earnings would be reduced to zero and not minus four.

Your net earnings after the second stage will be calculated as follows:

$(\text{Gross Earnings from Stage One}) - (2 \times \text{the Number of EMUs spent on reductions directed towards you}) - (\text{your expenditure on reductions directed at other members of the group})$ .

If you have any questions please raise your hand. Otherwise, click the red FINISHED button when you are done reading.

This is the end of the instructions. Be patient while everyone finishes reading.

c:\Papers\Archives\MutualMonitoring\Mutual.tex March 2, 2009

	(1)	(2)	(3)
$\sum_{j \neq i, t-1} \sigma_j / n$ (lag average contribution of teammates)	0.48 (0.06)***	0.44 (0.05)***	
$q$ (the productivity of the public good)	1.43 (0.36)***	0.80 (0.26)***	0.75 (0.25)***
$n$ (team size)	-0.60 (0.31)**	-0.31 (0.20)	-0.30 (0.20)
$(1 - \sigma_{i, t-1})w$ (lag contribution)		0.38 (0.04)***	0.82 (0.06)***
$\sum_{j \neq i, t-1} s_{ji}$ (lag punishment received)		0.13 (0.04)***	0.02 (0.09)
Shirker's Deviation			0.28 (0.10)***
Contributor's Deviation			-0.53 (0.09)***
Shirkers Deviation $\times$ Lag Punishment			0.03 (0.01)***
Contributor Deviation $\times$ Lag Punishment			-0.05 (0.02)***
Constant	6.93 (1.72)***	1.63 (1.14)	2.36 (1.44)
$\chi^2$	132	284	323
$N$	1548	1548	1548

**Table 2:** Testing the Determinants of Contributions. Note: all regressions are Tobits, include random effects and time period fixed effects. The dependent variable is one's contribution in round  $t$ . Standard errors of the estimates are in parentheses. \* indicates significant at the 0.10, \*\* 0.05, \*\*\* 0.01 levels.