Mutual Monitoring in Teams: Theory and Evidence on the Importance of Residual Claimancy and Reciprocity*

Samuel Bowles
Jeffery Carpenter
Herbert Gintis†

January 24, 2001

Abstract

Monitoring by peers in work teams, credit associations, partnerships, local commons situations, and residential neighborhoods is often an effective means of attenuating incentive problems. Most explanations of the efficacy of mutual monitoring rely either on small group size or on a version of the Folk theorem with repeated interactions. We provide a model of team production in which the effectiveness of mutual monitoring depends not on these factors, but rather on 'strong reciprocity': the willingness of some team members to engage in the costly punishment of shirkers.

By performing a laboratory experiment involving a public goods game with punishment, we provide empirical evidence for the behavioral relevance of strong reciprocity in teams, including the fact that the willingness to punish shirkers increases with the team's residual claim and does not decrease in larger teams. We conclude with some results specifying conditions under which mutual monitoring in teams provides an effective solution to incentive problems arising from incomplete contracting, as well as conditions under which mutual monitoring is likely to fail.

1 Introduction

Monitoring by peers in work teams, credit associations, partnerships, local commons situations, and residential neighborhoods is often an effective means of attenuating incentive problems that arise where individual actions affecting the well being of

*We would like to thank the MacArthur Foundation for financial support.

†Samuel Bowles and Herbert Gintis are at the Department of Economics, University of Massachusetts, Amherst, and Jeffery Carpenter is at Middlebury College, Middlebury Vermont.

others are not subject to enforceable contracts (Whyte 1955, Homans 1961, Ostrom 1990, Tilly 1981, Hossain 1988, Dong and Dow 1993b, Sampson, Raudenbush and Earls 1997). Most explanations of the incentives to engage in mutual monitoring (Varian 1990, Stiglitz 1993) rely either on the small size of the interacting group, or on repeated interactions and low discount rates, allowing the Folk theorem to be invoked. Neither of these is completely satisfactory, since work teams are often large and the Folk theorem has little explanatory power.[1] Other treatments leave the incentive to engage in mutual monitoring unexplained (Arnott 1991, Weissing and Ostrom 1991).[2]

We provide a model of team production in which the effectiveness of mutual monitoring depends not on these factors, but rather on 'strong reciprocity': the willingness of some team members to engage in the costly punishment of shirkers. The key conditions supporting mutual monitoring are (a) the fact that when members are residual claimants, shirking imposes costs on other team members, and contributing to production becomes a team norm, and (b) a fraction of team members are 'reciprocators' who punish violators of team norms.

By performing a laboratory experiment involving a public goods game with punishment, we provide empirical evidence for the behavioral relevance of strong reciprocity in teams, including the fact that the willingness to punish shirkers increases with the team's residual claim and does not decrease in larger teams. We conclude with some results specifying conditions under which mutual monitoring in teams provides an effective solution to incentive problems arising from incomplete contracting, as well as conditions under which mutual monitoring is likely to fail.

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 monitoring itself, while addressing the members' incentive to free ride by the threat of dismissal by the monitor. They contrast this view of the 'classical firm,' as they call it, with an alternative in which team members are residual claimants and monitor-

[1] The repeated game solution to the problem of sustaining cooperative behavior in teams has several weaknesses, including: (a) there are a multiplicity of equilibria, most of which do not exhibit high levels of cooperation; (b) subgame perfection (i.e., the credibility of threats to punish non-cooperators) requires an implausible degree of coordination among team members.

[2] Dong and Dow (1993b) and Legros and Matthews (1993) 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 (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.

ing 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 to team members. These arrangments 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.

These approaches do not explain how mutual monitoring works, but rather why it may be unnecessary. The limited applicability of the owner-monitor and optimal contracting approaches provides one motivation for exploring the relationship between residual claimancy and mutual monitoring 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 (Hansen 1997). Mutual monitoring 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 (Hossain 1988). Experimental studies (Frohlich, Godard, Oppenheimer and Starke 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 monitoring 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, Schlozman and Brady 1995), as well as willingness to monitor and 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 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 mutual monitoring.[3] Monitoring is costly, however, and if the desire to monitor is not sufficiently widespread, we shall see, mutual monitoring will fail.

2 Residual Claimancy and Strong Reciprocity

We will show that under certain conditions, residual claimancy by team members can provide sufficient incentives for mutual monitoring, and thus support high levels of team performance. A key element in our approach, one shared by recent contributions of Kandel and Lazear (1992), Rotemburg (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 under appropriate conditions the assignment of residual claimancy to team members will attenuate incentive problems even when team size is large.

Two common characteristics of successful mutual monitoring 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 monitoring in large groups.[4]

[3]Some models of mutual monitoring are presented in Varian (1990), Kandel and Lazear (1992), Weissing and Ostrom (1991), Dong and Dow (1993a,b), and Banerjee, Besley and Guinnane (1994). Other models of incentives in teams include Holmström (1982), McAfee and McMillan (1991), Legros and Matthews (1993), Rotemburg (1994), and Besley and Coate (1995)

[4]The problem of motivating the peer-monitors 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. Rotemburg (1994) develops a model of this type. More generally, Robert Frank writes: "Under [profit sharing] plans, the injury caused by an act of shirking affects not only the shareholders of the firm but also the shirker's co-workers. Individual workers who care about their co-workers will be reluctant to impose these costs…even when it is impossible for co-workers to observe the act of shirking." (1991):168. However were team members sufficiently

A parsimonious explanation of mutual monitoring is provided, however, by the notion of *strong reciprocity*: the well-documented human propensity to cooperate with those who obey, and to punish those who violate social norms, even when this behavior cannot be justified in terms of self-regarding, outcome-oriented preferences (Campbell 1983).[5] We distinguish this from *weak reciprocity*, namely reciprocal altruism, tit-for-tat, exchange under complete contracting, and other forms of mutually beneficial cooperation that can be accounted for in terms of self-regarding outcome-oriented preferences. The commonly observed rejection of substantial positive offers in experimental ultimatum games is consistent with this interpretation.[6] Moreover the fact that offers generated by a computer rather than another person are significantly less likely to be rejected suggests that those rejected offers at to cost to themselves are reacting to violations of norms rather than simply rejecting disadvantageous offers (Blount 1995). More directly analogous to the team production case, however, are findings in n-player public goods experiments. These provide a motivational foundation for mutual monitoring 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 a group to which they belong badly.[7] In these experiments, which allow subjects to punish non-cooperators at a cost to themselves, the moderate levels of contribution typically observed in early play tend to 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, punishment levels are undiminished in the final rounds, suggesting that disciplining norm violators is an end in itself and hence will be exhibited even when there is no prospect of modifying the subsequent behavior of the shirker or potential future shirkers.

The willingness to engage in costly punishment provides a basis for linking residual claimancy with mutual monitoring, even in large teams. An individual

altruistic in this sense to motivate mutual monitoring, there would be no initial free rider problem either.

[5] Kandel and Lazear (1992), which is otherwise closest to our approach to modelling mutual monitoring, do not admit a reciprocity motive, but rather assume that members monitor and punish to increase their individual material payoffs.

[6] See Güth, Schmittberger and Schwarz (1982), Güth and Tietz (1990), Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991), Ostrom, Walker and Gardner (1992), Güth and Ockenfels (1993), Forsythe, Horowitz, Savin and Sefton (1994), Camerer and Thaler (1995), Cameron (1995), Hoffman, McCabe and Smith (1998), and Falk and Fischbacher (1998). For an overview of the studies in this area, see Davis and Holt (1993) and Fehr, Gächter and Kirchsteiger (1997).

[7] See Ostrom et al. (1992) on common pool resources, Fehr et al. (1997) on efficiency wages, and Fehr and Gächter (2000) on public goods. Coleman (1988) develops the parallel point that free riding in social networks can be avoided if network members provide positive rewards for cooperating.

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. We should note that our model requires only that a certain 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 retaliate, either because they are self-interested, or they are purely altruistic.[8]

In support of the analytical model developed in Section 3 we report in Section 5 an experiment carried out by the authors involving a public goods game with costly punishment. This experiment replicates 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 repeated. In addition, we show that the level of punishment directed towards a team member increases with the social cost that member imposes on the group due to shirking. In particular, the level of punishment and cooperation increase with the degree of team residual claimancy, and do not decrease when team size increases. Moreover, players appear to understand quite clearly the incentives they create and to which they respond, since after the first few rounds of play, punishment is directed virtually exclusively towards shirkers, and punishment in one round leads shirkers to increase their contribution in the next round. Finally, we show that these results are not due to pure altruism on the part of punishers, since players punish shirkers even when the costs of punishing shirkers exceed the increase in group earnings afforded by punishing.

3 Mutual Monitoring in Teams

Consider a team with n members ($n > 3$), each of whom can either work, supplying one unit of effort, or shirk, supplying zero units of effort. We assume the members of the team are equal residual claimants on team output, but there may be other residual claimants outside the team (e.g., equity-holders who do not engage in production, or a government that taxes output). For convenience, if we refer to i or j, we assume they are team members in $\{1, \ldots, n\}$ unless otherwise stated, and if we refer to both i and j, we assume $i \neq j$. Also we write $n_{-i} = \{k = 1, \ldots, n | k \neq i\}$. We assume agents have linear utility functions that are additive in costs and benefits.

Let σ_j be the probability that member j shirks, so $\sigma = \sum_{j=1}^{n} \sigma_j / n$ is the average rate of shirking. We assume the cost of working (not shirking) is one dollar, and working adds q dollars to team output. Each member's payoff is then given by $\alpha q(1-\sigma)$, where $\alpha \in [0, 1]$ is defined as the team's residual share. The sum of the

[8]For an especially clear example, see Blount (1995). Fehr and Schmidt (1999) provides a survey of rejection rates in ultimatum games.

payoff losses to the team from one member shirking is αq, of which the shirker's share is $\alpha q/n$, so the shirker's net gain from shirking is $g = 1 - \alpha q/n$, which we assume is strictly positive. We also assume $\alpha q > 1$, otherwise universal shirking would be optimal.[9]

Consider a single team member j. Another member $i \in n_{-j}$ can either monitor j at cost $c_i > 0$, or trust j at zero cost. Member i imposes a cost $s_j > 0$ on j if i detects j shirking. This cost may involve public criticism, shunning, threats of physical harm and the like. We assume that acts of punishment, like work effort, are non-verifiable and hence not subject to contract. If j shirks and i monitors j, we assume the shirking will be detected with probability $p_{ij} \in (0, 1]$, where this *probability of detection* may vary with the spatial proximity of team members and the transparency of the production process, and other factors that we do not model here.

To model the incentive to punish, let z be the sum of the payoffs lost to the $n - 1$other team members when one team member shirks with probability one, so $z = \alpha q(n-1)/n$ and $z\sigma_j$ is the expected loss imposed on other members of the group when j shirks with probability σ_j. We refer to z as the *social cost of shirking*, and we suppose that i experiences a subjective gain $\rho_i z\sigma_j$ from disciplining a shirking member j, which occurs if j shirks, i monitors j, and j is detected. We assume $\rho_i \geq 0$.[10] We call ρ_i the *propensity to punish shirkers* of member i.[11] Note that some members may exhibit no propensity to punish; i.e., $\rho_i = 0$.

If i monitors, the likelihood of detecting j shirking is $\sigma_j p_{ij}$, so the net cost of monitoring j over trusting j is $c_i - p_{ij}\rho_i z\sigma_j$. Then if μ_{ij}, $i \in n_{-j}$, the probability that i monitors j, is chosen to be a best response, we have

$$\mu_{ij} \begin{cases} = 0, & c_i > p_{ij}\rho_i z\sigma_j \\ \in (0,1), & c_i = p_{ij}\rho_i z\sigma_j \\ = 1, & c_i < p_{ij}\rho_i z\sigma_j \end{cases} \tag{1}$$

[9] Most of the homogeneity assumptions we make can be dropped, at the expense of complicating the notation and the descriptions of the model.

[10] Note that, unlike a member's share of the firm's net revenue, the subjective gain from punishing does not decline with the size of the team. We motivate this assumption and discuss the effects of team size below. We assume $\rho_i z < s_j$ to avoid bizarre '*sado-masochistic*' equilibria in which team members always shirk and always punish shirkers, the net psychic return to which is greater than the return to working.

[11] For simplicity we have assumed that a monitor's propensity to punish, ρ_i, is not affected by the propensities to punish or the observed rates of punishing of other members of the team. Replacing this with the assumption that punishing propensities are positively related opens the possibility of multiple equilibria, some involving high levels of punishing and some low. We explore this alternative below.

Let s_j^* be the expected punishment inflicted by all $i \in n_{-j}$ on j if j shirks. We have

$$s_j^* = \sum_{i \in n_{-j}} p_{ij}\mu_{ij}s_i. \tag{2}$$

Writing the direct gain to an individual from shirking as

$$g = 1 - \alpha q/n, \tag{3}$$

the expected gain to j from shirking, including the expected cost of punishment, is $g - s_j^*$. Therefore if σ_j is chosen as a best response, we have

$$\sigma_j \begin{cases} = 0, & g < s_j^* \\ \in [0, 1], & g = s_j^* \\ = 1, & g > s_j^* \end{cases} \tag{4}$$

If $g < 0$ there is a unique, Pareto efficient, Nash equilibrium in which no members shirk and no member monitors.[12] In this case residual claimancy alone is sufficient to ensure efficiency. The more interesting case, however, is where group size is sufficiently large that residual claimancy alone does not entail incentive compatibility. We thus suppose in the rest of the paper that $\alpha q/n < 1$, so that even with full residual claimancy assigned to the team as a whole, shirking is an individual best response in the absence of monitoring. In this case any Nash equilibrium involves positive shirking, since if $\sigma_j = 0$ for some j then by (1) $\mu_{ij} = 0$ for $i \in n_{-j}$. But then by (4), $\sigma_j = 1$, a contradiction. Thus we must investigate conditions under which $0 < \sigma_j < 1$ for some j in equilibrium, requiring

$$g = s_j^*. \tag{5}$$

We call such a situation a *working equilibrium*.

We say i is a *reciprocator* if $\rho_i > 0$. Suppose the fraction of reciprocators is f, so the remaining fraction $1 - f$ of team members are self-regarding—for these agents $\rho_i = 0$, and they never monitor or punish. Notice that if j is not a reciprocator, j has fn potential monitors, whereas if j is a reciprocator, j has $fn - 1$ potential monitors. In the a interest of simplicity of exposition, we will ignore this difference, assuming all agents face fn potential monitors.[13]

We say reciprocators are *homogeneous* if there are *parameters* p, c, s and ρ such that if i is a reciprocator, then $c_i = c$, $p_{ij} = p$, $s_i = s$, and $\rho_i = \rho$. We have

[12] This equilibrium is efficient because we have ruled out 'sado-masochistic' equilibria.

[13] The effect of dropping this assumption is in all cases quite transparent, since in effect, the model is the union of n independent games, in each of which one agent is the worker and the other $n - 1$ agents are the monitors.

Theorem 1. *Suppose reciprocators are homogeneous, with parameters p, c, s and ρ. If $c > p\rho z$, then the unique Nash equilibrium satisfies $\sigma_j = \sigma^* = 1$ and $\mu_{ij} = \mu^* = 0$ for all $i, j = 1, \ldots, n$; i.e., all members shirk and no member monitors. If $c < p\rho z$ and $g > fnps$ the unique Nash equilibrium satisfies $\sigma_j = \sigma^* = \mu^* = 1$ for all $i, j = 1, \ldots, n$; i.e., all workers shirk and all members monitor. If $c < p\rho z$ and $g < fnps$, then*

(a) there is a mixed strategy Nash equilibrium in which all members shirk with probability

$$\sigma^* = \frac{c}{p\rho z} \tag{6}$$

and reciprocators monitor with probability

$$\mu^* = \frac{g}{fnps}; \tag{7}$$

(b) this equilibrium remains when team size increases;

(c) increased residual claimancy reduces the incidence of shirking. This effect does not decrease when team size increases.

(d) the social welfare difference per team member between a first best world with no shirking and the equilibrium of this game is $\sigma^*(\alpha q - 1)$. This is increasing in the degree of residual claimancy and does not decline when team size increases.

Proof: If $c > p\rho z$, then the payoff to i from monitoring j is negative even when $\sigma_j = 1$. Hence $\mu_{ij} = 0$. But then $s_j^* = 0$, and since $g > 0$ by (4), we have $\sigma_j = 1$. If $c < p\rho z$ and $g > fnps$, then from (2), $s_j^* = fnp\mu s < g$ for any j, so again (4) shows $\sigma_j = 1$. Then from (1) we have $\mu_{ij} = 1$ whenever i is a reciprocator.

If $c < p\rho z$ and $g < fnps$, it is easy to check that there is no pure strategy Nash equilibrium. By our homogeneity assumption, in a mixed strategy Nash equilibrium, all members will choose the same rate of shirking σ^* and reciprocators will choose the same rate of monitoring μ^*. But (σ^*, μ^*) is a Nash equilibrium when the payoffs to Trust and Monitor are equal and the payoffs to Work and Shirk are equal. From (1), the payoffs to Trust and Monitor are equal when $c = \sigma^* p\rho z$, which proves (6). Similarly, we see from (2) and (4) that the payoffs to Work and Shirk are equal when $g = f\mu^* nps$, which proves (7).

Part (b) holds because the inequalities on which the existence of this equilibrium depends continue to hold when team size is increased. Part (c) is true by (6) and because the propensity of punish defectors is rising in the residual share. The result that both σ^* and μ^* are independent of group size and the frequency of reciprocators holds because the net benefits of membership or trusting (which σ^* equates) are also independent of n and f. For part (d), note that With no shirking the net social surplus per team member is just $\alpha q - 1$. When all members shirk with probability σ, social surplus is reduced by $\sigma(\alpha q - 1)$ per member. In their roles as monitors, Trusting incurs no costs and produces no benefits in equilibrium, and the net benefits of monitoring and trusting are equal so members experience no net gains from monitoring. The net surplus of the team per member is thus $(1 - \sigma)(\alpha q - 1)$, from which the result follows. ■

While the fraction of reciprocators and the level of punishment they may inflict do not appear in (6), they are not unimportant in the determination of the level of shirking, for if f and s are sufficiently small, the condition $g < fnps$ is violated, and universal shirking occurs.[14] Notice also that, as one would expect, shirking declines with an increase in the propensity to punish shirkers, an increase in the probability of detecting shirking, or a decrease in the cost of monitoring.

The intuition behind part (a) of Theorem 1 is as follows. The equilibrium level of shirking, σ^*, equates the net benefits of monitoring and trusting, while the equilibrium level of monitoring equates the net benefits of working and shirking. Thus when σ is greater than its equilibrium value (6), the expected benefits of monitoring, $\sigma p\rho z$, exceed the costs c. Members who monitor with high probability will then receive higher payoffs than members who monitor with low probability, inducing some to increase their monitoring probability. As the monitoring probability increases, the gains to shirking decline, leading suppliers to reduce σ. This dynamic continues until (6) is satisfied. A similar dynamic occurs when σ is less than its equilibrium value. When μ is greater than its equilibrium value (7), the expected costs of shirking $f\mu nps$ exceed the benefits g. Suppliers who work with high probability will then be receiving higher payoffs than suppliers who shirk with high probability, inducing some to decrease their rate of shirking. As the shirking rate declines, the gains to monitoring decline, leading to a reduction in μ. This dynamic will continue until (7) is satisfied. A similar dynamic occurs when μ is less than the equilibrium value given by (7).

Behavioral traits such as a work ethic or a willingness to punish co-members for inflicting harm on the team are, of course, strongly norm=governed and as such need not be proximately determined by the explicit optimization of any agent but rather may be the expression of behavioral rules. Thus the model underlying Theorem 1 may be interpreted as the basis of a dynamic treatment of work and punishment norms, with the updating of norms responding to the observed payoffs of others. For example, as our description of the intuition behind part (a) of Theorem 1 suggests,

[14] We have not investigated equilibria when $c = p\rho z$ or $g = fnps$, since these cases are neither generic nor interesting.

the determination of σ and μ may be represented as dynamic processes based on the differential replication of norms governing the working, shirking, monitoring and punishing behaviors we have modeled, the equilibrium values σ^* and μ^* simply representing outcomes that are stationary in the underlying dynamic. We do not develop this extension here.

Relaxing the assumption that reciprocators are identical, we say reciprocators are *quasi-homogeneous*, if there are parameters p, c, and s such that if i is a reciprocator, then $c_i = c$, $p_{ij} = p$, $s_i = s$. We have

Theorem 2. *Suppose reciprocators are quasi-homogeneous with parameters p, c, and s. We order the team members so that*

$$\rho_1 \geq \ldots \geq \rho_n,$$

and let k be the unique integer that satisfies[15]

$$k < \frac{g}{ps} < k + 1, \tag{8}$$

Then if $c > p\rho_1 z$ there is a unique Nash equilibrium with $\sigma^* = 1$ and $\mu^* = 0$ (all members shirk and no member monitors). Suppose that $c < p\rho_1 z$. Then if $k \geq n$, there is a unique Nash equilibrium in which $\sigma^* = \mu^* = 1$ (all agents shirk and all agents monitor). Otherwise there is an equilibrium such that agents $1, \ldots, k$ monitor with certainty, agents $k + 2, \ldots n$ do not monitor, agent $k + 1$ monitors with probability

$$\mu_{k+1} = \frac{g - kps}{ps}, \tag{9}$$

and the rate of shirking is given by

$$\sigma^* = \frac{c}{p\rho_{k+1} z} < 1. \tag{10}$$

If $\rho_{k+1} \neq \rho_k$, this equilibrium is unique.

Finally, suppose

$$p\rho_{l+1} < c < p\rho_l \tag{11}$$

for some $l < n$. Then if $k \geq l$ there is a unique Nash equilibrium in which $\sigma^* = 1$ and members $1, \ldots k$ monitor with certainty while members $k + 1, \ldots n$ trust. If $k < l$, there is a unique Nash equilibrium in which σ^* is given by (10), members $1, \ldots k$ monitor with certainty, member $k + 1$ monitors with probability $g/ps - k$, and no other members monitor. This exhausts the (generic) cases.

[15]We assume the generic case where g/ps is not an integer.

Proof: If $c > p\rho_1 z$ then $c > p\rho_i z$ for all i, so all members trust and hence all members shirk. Suppose $c < p\rho_n z$. If $k \geq n$ then $g > nps$ so members shirk even if every member monitors, which occurs because $p\rho_i z > c$. If $l \leq k < n$, then with $\sigma^* = 1$, members $1, \ldots l$ monitor and other members trust, so $s^* = lps \leq kps < g$, confirming that $s^* = 1$ is a best response. finally, if (11) holds and μ_{k+1} is given by (9), then $g = s^*$, so the net benefits of shirking and working are identical for all members. If σ is given by (10), agent $k+1$ is indifferent between monitoring and trusting. Thus the given values of σ and μ_{k+1} are best responses. We also know that since member $k > 1$ monitors, then members $1, \ldots, k$ monitor, since $c \leq \sigma p\rho_k z$ implies $c \leq \sigma p\rho_i z$ for $i = 1, \ldots, k$. Therefore this is a Nash equilibrium in which the first $k+1$ members monitor. If $\rho_{k+1} > \rho_k$ then this equilibrium is unique, since members $1, \ldots, k$ must now monitor with probability one. ■

Theorem 2 shows that the preferences of marginal agent $k+1$ determine the level of shirking (10) for the team as a whole, while the level of monitoring necessary to sustain σ^* as an equilibrium is, by (9), determined by the location of the margin (that is by k). Of course team members may differ with respect to not only their propensity to punish, but their marginal disutility of labor, the cost of monitoring, the ability of each monitor to detect shirking, and the amount of punishment they may inflict as well. Generalizing the sources of heterogeneity among team members is readily accomplished.

4 Impediments to Mutual Monitoring: Team Size and ‘Shirking Cliques’

Would it be plausible to treat ρ as a decreasing function of team size, on the grounds that strong reciprocity may weaken when the team becomes larger, and thus the propensity to punish any given act of shirking would fall? Though this is possible, there is to our knowledge no clear evidence in support of this notion, and there are many ‘stylized facts’ contradicting it. For instance, citizens of large nations appear no less willing to sacrifice for their compatriots than those of small nations. Similarly, people are often observed to support their local sports team, their regional sports team, and their national sports team with equal commitment. Examples of this type abound. The experimental evidence described below, moreover, suggests that ρ does not depend on team size.

There are of course additional paths through which increasing team size might weaken the mutual monitoring mechanism. Increased n might lower the average cost s a monitor can impose on a shirker, since the ‘average social distance’ between a pair of workers can be expected to increase as the team becomes more numerous. The ability to detect shirking among randomly selected pairs of team members may also decline. Notice, however, that adding team members also increases the number of potential monitors so it might well increase the likelihood that any given act of shirking would be detected, and hence might increase the amount of punishment which a shirker would expect.

We say *i sees j* if $p_{ij} > 0$. To this point we have assumed that for all i, j, i sees j. We now weaken this assumption, getting

Theorem 3. *Suppose there is an integer $\kappa \geq 1$ such that as team size $n \geq \kappa$ increases, each member sees exactly κ other team members. Then if $f\kappa ps > g$, the assertions of Theorem 1 hold, with κ substituted for n and with the level of shirking in (6) unaffected.*

The intuition behind Theorem 3 is that as long as increasing team size does not reduce the number of team members that one may see the effectiveness of mutual monitoring does not decline as team size increases. The proof is as in Theorem 1, with minor changes. Notice, however, that the effectiveness of mutual monitoring depends on the extent to which members may be informed about one another's activities: if team members are able to see relatively few co-members, then the condition $f\kappa ps > g$ is not likely to obtain, and universal shirking will be the equilibrium.

However large teams often do not have the informational homogeneity assumed in Theorem 3, since with increased team size often comes a more refined division of labor in which there are specialized 'work groups' whose members all see one another, and who are not seen by other team members. If members of such a group have an incentive to make credible commitments involving the reciprocators in the group not monitoring, so all members of the group can shirk without penalty, we call the group a *shirking clique*. We have

Corollary 3.1. *Suppose the conditions of Theorem 1 hold, except that there is an integer $\kappa \geq 1$ such that each member sees exactly κ other team members. Suppose there is a mixed strategy equilibrium defined by (6) and (7), except $p = 0$ when a member does not see another member. Finally, suppose there is a group C that is isolated—members of C see one another but are not seen by non-C members. Then if the frequency of reciprocators in C is f (the same as in the team), C is not a shirking clique.*

Proof: This follows from the observation that Theorem 3 holds, and this theorem assumes nothing concerning the pattern of seeing. The point is that a group that sees only its own members cannot agree that the reciprocators will not monitor and punish, since there is no way to enforce such an agreement within the model and reciprocators would benefit from violating the agreement. ■

Yet shirking cliques may exist under less restrictive conditions. We will mention, but not develop formally, three possibilities. First, if group composition is not random with respect to the frequency of reciprocators, it is clear from Theorem 1 that an isolated group may have a sufficiently low frequency of reciprocators that it will form a shirking clique. The critical frequency of reciprocators f_c for an isolated group of size k is

$$f_c = \frac{g}{kps},$$

indicating that shirking cliques may be more difficult to sustain in larger groups.

Second, we have assumed that the degree of strong reciprocity depends on the team's residual share, which is a fraction α of the net payoff to production. But just as a team member does not care about the payoff to non-team members, the team member may not care about a subset of fellow team members. Suppose there is a subgroup G of team members whose degree of strong reciprocity depends only on the residual share received by G. We call G a *group of insiders*.[16] Our assumption to this point is that the whole team is a group of insiders. But suppose there is some social distinction (race, ethnicity, gender, age, creed, and the like) that gives rise to an isolated group G of insiders that comprise a fraction $v < 1$, of the team. Strong reciprocity for members of G depends then on $v\alpha < \alpha$, and for sufficiently small v, by Theorem 1, G will sustain a shirking equilibrium, making the group a shirking clique. If the remainder of the team were also a group of insiders, their incentive to monitor would be based on a subjective residual claimancy of $(1 - v)\alpha$, which for sufficiently small v would be sufficient to sustain a working equilibrium among non-G team members. Notice also that if there is sufficient social heterogeneity that many insider groups exist within the team, a working equilibrium for the team will not exist.

A third type of shirking clique emerges if we admit asymmetric information and collusion. Suppose a work group is isolated and collude to impose costs on reciprocators, making it unattractive to monitor. Then the group can form a shirking clique. Even a non-isolated group whose members collude to punish reciprocators may constitute a shirking clique as long as there are not too many reciprocators among non-clique members who can see the members of the group. Indeed, our experimental data indicate that shirkers do tend to punish reciprocators.

To model this case, consider a group C of κ members in an n-member team satisfying the conditions of Theorem 1, and which is in a mixed strategy Nash equilibrium defined by (6) and (7). We alter the informational assumptions of the model so that collusion cannot be detected, in the following sense: if the members

[16] On the general tendency for people to categorize others as 'insiders' and 'outsiders,' and behave differentially towards the two types, see Krebs and Denton (1997) and the references therein.

of C collude to shirk, this is common knowledge among the members of C, but nonmembers of C continue to believe that members of C shirk at rate σ^* given by (6), and hence continue to monitor members of C at rate μ^*. Suppose also that members of C can enforce a promise on the part of a reciprocator in C not to monitor others in C. We say the model *supports collusion* if these two conditions hold.

Suppose the conditions of Theorem 1 hold and there is a mixed strategy Nash equilibrium defined by (6) and (7). Consider a group C with κ members, the frequency of reciprocators within which is f_C. If the model supports collusion, then C is a shirking clique exactly when

$$f_C \geq \frac{(1-\sigma)\alpha w f}{g}. \tag{12}$$

To see this, note that the gain from shirking for a member of C is g. The cost s_c^* of shirking is $\mu^*(fn - f_C\kappa)ps = g - gf_C\kappa/fn < g$, so members of C will shirk, and each will gain $gf_C\kappa/fn$. But by shirking C imposes a loss of $(1-\sigma)\alpha w\kappa$ on the group, the cost of which to each member of C is $(1-\sigma)\alpha w\kappa/n$. Thus members of C will gain from collusion exactly when $(1-\sigma)\alpha w\kappa/n < gf_C\kappa/fn$, which simplifies to (12).

5 Experiments in Mutual Monitoring

Our model of mutual monitoring in teams depends critically on the underlying behavioral assumption, namely that under some conditions strong reciprocity motives will induce sufficient punishment levels to sustain high levels of team output. We ran an experimental public goods game to test the plausibility of this assumption, 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 others 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 punishment setting prevented the decay of cooperation usually found in public goods experiments.[17]

In addition to replicating Fehr and Gächter (2000), we investigate the effect of the degree of residual claimancy and group size on the propensity to punish.[18] Whereas Fehr and Gächter frame their experiment as a "group project," we use the term "voluntary contribution to an investment," since this more neutral language leads to a more stringent test of our hypotheses. Moreover we implement only one

[17] For a review of the results of the standard public goods game, see Ledyard (1995).

[18] The instructions for a typical session appear in the Appendix. A more complete description of the experimental design is available from the authors.

of their treatments, the so-called Stranger Treatment in which subjects are randomly reassigned to a new group after each round of play. This also adds to the stringency of our experiment, since the more standard Partner Treatment, in which teams remain together throughout the experiment, tends to foster more cooperation than the Stranger Treatment (Croson 1996, Fehr and Gächter 2000). Finally, we make punishing shirkers quite costly to punishers, the cost of inflicting one unit of penalty on a low contributor rising from a unit's loss to the punisher at low levels to three units at high levels. Implementing a very high cost of punishing is useful, since it biases the experiment against our hypotheses, but it is probably quite unrealistic in many social settings, where punishing costs are generally quite low (Bingham 1999, Gintis 2000).

Suppose there are n players, each player receives w 'francs' (exchangeable for dollars at the end of the experimental session) 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 by assigning from zero to ten 'points' to particular players.[19] Let l_{ij} be the number of points assigned by player i to player j (we assume $l_{ii} = 0$). Then the payoff to player i is

$$\pi_i = w\left[\sigma_i + nm(1-\sigma)\right](1 - \min[1, l_i]) - \sum_{j=1}^{n} c(l_{ij}), \tag{13}$$

where $\sigma = \sum_{j=1}^{n} \sigma_j/n$ is the average shirking rate, $l_i = \sum_j l_{ji}/10$ is the average number of points assigned to i, and $c(l_{ij})$ is the cost to i of imposing l_{ij} points on j. Note that we have set $\alpha = 1$ and defined a new variable $m = q/n$, since we do not use the concept of residual claimancy in the experiment, but rather vary m, which is the per-member payoff to a player contribution of one dollar, and n, the team size. Also, one dollar in the model developed in the previous two sections equals $w = 20$ francs in the experiment. The cost of given points was set to

l_{ij}	0	1	2	3	4	5	6	7	8	9	10
$c(l_{ij})$	0	1	2	4	6	9	12	16	20	25	30

To study the effect of team size, one treatment used teams of five subjects, and another treatment increased the team size to ten. Summing the payoffs to all team members given by (13), we see that z, the harm per unit of shirking inflicted on the members of the team other than the shirker, is equal to $z = wm(n-1)$. We set z to 24, 54, 60, and 135, corresponding to $m = 0.3$ in two groups (one large and one

[19] The instructions to participants refer to "assigning points" with no interpretation supplied. In early periods the correlation between the contribution of and points assigned to a player is virtually zero, but by period ten points are assigned overwhelmingly to low contributors.

small) and $m = 0.75$ in the other two groups, given that $w = 20$. A total of twelve sessions were conducted with 205 participants. Figure 1 summarizes the design and the session numbers associated with each treatment.[20] The number of participants, and therefore teams, per treatment vary due to no-shows. All subjects were recruited by email from the general student population 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 computer terminal which was isolated (there were blinds on each side) so that decisions were made in privacy. Each session took approximately 45 minutes from sign-in to payments and subjects earned $19.81 on average, including the show-up fee.

	five person team	ten person team
$m = 0.30$	7 teams 35 subjects $z = 24$	4 teams 40 subjects $z = 54$
$m = 0.75$	6 teams 30 subjects $z = 60$	4 teams 40 subjects $z = 135$

Figure 1: Experimental Design

Each session lasted ten periods. In each period (a) subjects were given the endowment of 20 $w = 20$ francs and allowed to contribute, anonymously, a 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 then revealed to each subject, who were then permitted to assign points to others. Finally, payoffs were calculated according to (13), and subjects were informed of their net payoffs for that period. Except in the final period, subjects were then randomly reassigned to a new group, and the process repeated.

If we assume standard 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 can have no effect on future payoffs to the punisher, given the Stranger

[20] In addition to the 145 subjects in the treatment groups, 60 subjects were assigned to no-punishment control sessions, with both low and with high m, to ensure that there is nothing unusual about our subject pool or procedures. We found that the control groups replicate the usual level and rate of decay of cooperation, except that in the small m treatment, the decay was less pronounced that in other studied, falling from 50% in round 1 to 38% in round 10. By comparison Croson (1996), with an $m = 0.5$ reports initial contributions of slightly more than 40%, falling to approximately 12% by the end of ten rounds. Fehr and Gächter (2000) for whom $m = 0.40$, find contributions in their strangers treatment of approximately 33% in the first period, declining to approximately 12% in period ten.

Treatment. Hence the unique subgame perfect Nash equilibrium of the game is that no one punishes and hence no player contributes to the public good. We want to test the following hypotheses. We use the fact that the social cost of player i's shirking is $z\sigma_i = wm(n-1)\sigma_i$ where m is the marginal per capita return, n is the group size, w is player i's endowment, and σ_i is player i's shirking rate. Equivalently, $z\sigma_i$ is the harm inflicted on other members of the group by player i's withholding some amount of the initial endowment rather than contributing it to the public account. We also define a 'shirker' on a given round to be a player who contributed less than average on that round.

Hypothesis 1: Punishing Occurs Whenever Shirking Occurs. There is a positive level of punishment in all periods under all treatment conditions.

Hypothesis 2: The Level of Punishment Increases with the Level of Shirking and the Harm that Shirking Inflicts on Other Team Members. The level of punishment directed toward player i increases with $z\sigma_i = wm(n-1)\sigma_i$, and indeed in m, n, and σ_i separately.

Hypothesis 3: Shirkers Respond to Punishment. Punishment in one round leads shirkers to increase their contribution in the next round.

Hypothesis 4: Punishment Fosters Cooperation. The level of cooperation does not decay when costly punishment is permitted.

Hypothesis 5: Altruism Does Not Explain Punishment. Players punish shirkers even when punishment does not lead to higher group earnings.

Hypotheses 1 and 4 replicate Fehr and Gächter (2000). Hypothesis 2 supports the assertions of the previous section, since it implies that the tendency to punish increases with the group's residual claim, and does not decline with increasing group size.

We measure the propensity to punish as the average number of points assigned per unit of shirking, defined as contributing one franc to one's private account. Concerning Hypothesis 1, Figure 2 shows that the propensity to punish is positive in all treatment conditions and does not decline in later periods except in the small size, low m treatment. Moreover, a majority of subjects punish. Pooling across treatments, we find that approximately sixty percent of participants are willing to pay a cost to reduce the earnings of another shirking team member.

Hypothesis 2 states that the propensity to punish is increasing in the social cost of shirking, and in each of the three components of the costs of shirking. Figure 3 show the results of a regression analysis of individual punishment decisions, where the dependent variable is the number of points that a participant received from other team members and the independent variable is the social cost of shirking, $z_i = z\sigma_i$. We see that z_i is significant at the 1% level in all periods, and the effect increases in later periods. Pooling all periods, we find that the social cost of shirking explains thirty percent of the variance in punishment and a one standard deviation increase

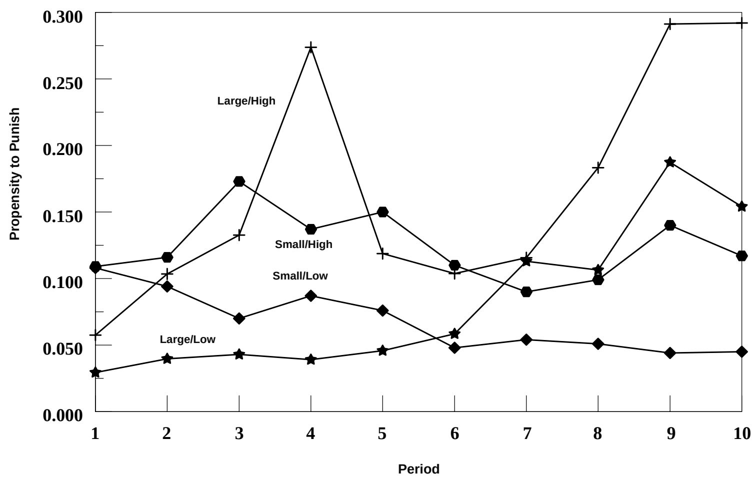

Figure 2: Propensity to punish, measured as average points given per act of shirking.

in z_i increases the punishment received by a shirker by 1.94 points (0.55 standard deviations).

It is illuminating to compare the regression results in Figure 2 across individual periods. First, the intercept diminishes over time and loses significance. Second, the coefficient on the social cost of shirking, $z_i = z\sigma_i$, increases both in magnitude and significance in later periods. Third, z_i explains an increasing fraction of the variation in punishment in later periods. These observations suggest a learning process whereby subjects become more consistently focused on using punishment as a disciplinary device as they gain experience with the game, and use the social cost of shirking as a measure of individual shirking.

Proposition 2 also suggests that the propensity to punish is increasing in each of the components of the social cost of shirking. This means the propensity to punish should be greater for the high m treatment than for the low, greater for the large team than the small, and should be directed toward subjects who shirk more rather than less. To test the effect of the three components of z_i on the propensity to punish, we regress the points received by participants on their private investment, a dummy variable that is unity for the high m treatment and a dummy for team size that is unity for large teams. The results of this test appear in Figure 4.

We also find that the relationship between the the propensity to punish and

	Dependant Variable = $\sum l_{ij}$ (Points assigned to player i by team members) (all results are OLS)										
	Period 1	2	3	4	5	6	7	8	9	10	All
intercept	2.75 (7.35)	2.26 (5.59)	1.77 (4.95)	1.23 (3.44)	1.25 (4.83)	0.69 (3.07)	0.86 (3.13)	0.56 (2.33)	0.27 (1.35)	0.42 (1.83)	1.21 (12.73)
$m(n-1)$ $\times w(1-\sigma_i)$	0.03 (3.20)	0.06 (5.76)	0.08 (6.54)	0.12 (7.82)	0.09 (9.11)	0.10 (11.22)	0.10 (7.51)	0.13 (9.81)	0.18 (13.02)	0.17 (15.58)	0.09 (25.20)
Adjusted R^2	0.07	0.19	0.23	0.30	0.37	0.47	0.28	0.40	0.54	0.63	0.30
F Statistic*	10 $p < .01$	3 $p < .01$	43 $p < .01$	61 $p < .01$	83 $p < .01$	126 $p < .01$	56 $p < .01$	96 $p < .01$	170 $p < .01$	243 $p < .01$	635 $p < .01$

Figure 3: Punishment and the Social Cost of Shirking.
Here σ_i is the private allocation of subject i, $m(n-1) = 1.2, 2.7, 3.0,$ or 6.75. The t-statistics are in parentheses.
*The F-statistic is on 143 degrees of freedom for each period, and 1448 degrees of freedom for All Periods.

team size behaves as posited by our model. Pooling the data for all periods, we find that the propensity to punish is greater in the high m treatment, $p < .01$, and higher in larger teams, $p < .03$ using both Kolmogorov-Smirnov (ks) tests for distributional differences and Wilcoxon tests for mean differences.[21] For the low m treatment the pooled propensities are distributed differently ($ks = .2971$, $p < .01$), largely because the propensity increases over time in large groups and declines or remains stable in small groups. Given the trajectories are nearly opposite the pooled average is not significantly different, however the ending propensities are significantly different ($z = 22.03$, $p < .01$). For the high m treatment, again, the propensities are distributed differently ($ks = .1889, p < .01$) and the pooled average is higher in the large group treatment, ($z = 12.14$, $p < .01$). Hence, the propensity to punish increases in larger teams.

To test the effect of team size on the propensity to punish, note that the marginal effect of a franc allocated to the private account on z_i is $m(n-1)$. In the low m, small team treatment this incentive is 1.20, for the low m, large team treatment $m(n-1)$ equals 2.70, for the high m, small team treatment $m(n-1)$ equals 3.00, and in the high m, large team $m(n-1)$ equals 6.75. Notice the incentive to punish is very similar in the low/large treatment and the high/small treatment (2.70 versus 3.00), with the incentive slightly lower in the larger group. If group size *per se* discourages punishment, we should find less punishment in the large team, low m treatment. But we do not. Comparing the propensities to punish graphed in

[21]Where appropriate, we test for distributional differences on the data pooled across all ten periods in additional to means tests to capture any effects due to differences in the time paths not accounted for in the Wilcoxon means test.

Figure 2, we find that large teams *punish more* in terms of both the distribution of points given to shirkers and the average level of punishment (ks= .2481, $p < .01$, $z = 18.64$, $p < .01$).

The dependent variable is the number of points assigned to player i by team members	
intercept	-0.38 (-1.96)
private	0.31 (23.41)
m-dummy	1.57 (9.94)
n-dummy	0.68 (3.83)
Adjusted R^2	0.31
F-Statistic*	2.15 $p < 0.01$

Figure 4: Punishers Punish Shirkers.
The results are for all periods pooled. The regression is OLS. Here 'private' is the private allocation of subject i, 'm-dummy' is unity for high m teams, and 'n-dummy' is unity for large teams. The t-statistics are in parentheses.
*The F-statistic is on 1446 degrees of freedom.

Hypothesis 3 asserts that being punished leads shirkers, whom we define as those subjects who contribute less than the average level for that round, to increase their next-round contribution. We test this by regressing the change in a shirker's public contribution between periods $t-1$ and t on the punishment points received by the shirker in period $t-1$ as a result of his or her contribution. The results are shown in Figure 5. We see that the coefficient on the points received last period term is positive and significant. One punishment point assigned to a shirker results in an increase of 0.38 of a point in subsequent contribution.

Hypothesis 4 asserts that the level of cooperation does not decay when costly punishment is permitted. The plausibility of this assertion follows from the level and efficacy of punishment exhibited above. We can test this directly, however, by plotting the average fraction of the endowment contributed to the public account against time, disaggregated by treatment. The results are shown in Figure 6. Note

The dependent variable is $x_{it} - x_{i,t-1}$, the difference in public contribution between $t-1$ and t	
intercept	1.11
	(3.13)
LagPts	0.38
	(6.50)
Adjusted R^2	0.08
F-Statistic*	42
	$p < 0.01$

Figure 5: Shirkers Respond to Punishment.
LagPts represents the points received in period $t-1$. The regression is OLS. The t-statistics are in parentheses.
*The F-statistic is on 468 degrees of freedom.

that contributions approach 100% in large teams, a result not reported previously in the experimental literature.

Finally Hypothesis 5 asserts that players punish shirkers whether or not punishment leads to higher group earnings. In other words, punishment cannot be accounted for as an instrumental strategy of altruists who want to increase the payoffs to team members. Figure 5 shows that, on average, shirkers increase their public contribution by .38 for each point received. Each increase by the shirker confers a benefit of $.38m(n-1)$ on the $n-1$ other members of the shirker's team next period, but comes at a cost to the punisher that is not less than 1 franc. Hence, in the Low/Small treatment the maximum expected benefit to the team of punishing (assuming the punisher altruistically evaluates gains and losses independently of which members they fall upon) is $.38(1.2) - 1 = -.54$. In the High/Small the benefit is .14, in the Low/Large .03, and in the High/Large it is 1.57. So even if the punisher values the (unknown) subjects who will be in the shirkers group next period as much as he values his own well being and even if he supposes all of them except the shirker are reciprocators, or at least worthy of his generosity, he would not punish in the Small/Low treatment if his reasons for punishment were instrumental (albeit altruistic). But significant levels of punishment occur in the small size, low m treatment, even in later rounds, as seen in Figure 2. We conclude that the motivation for punishment must include the desire to inflict a cost on shirkers.

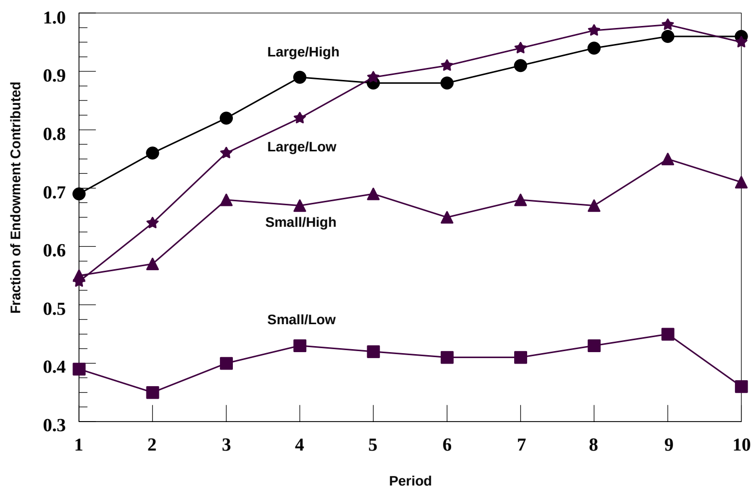

Figure 6: Average Contribution to Public Account

6 Conclusion

Our model and experimental results suggest that under appropriate conditions, strong reciprocity can support mutual monitoring even in large teams, unless the frequency of reciprocators is too low or the division of labor favors the formation of shirking cliques. Furthermore, mutual monitoring allows levels of member effort that are closer to first best, thus enhancing team welfare.

The case we have modeled appears to describe 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 opportunities to punish members who impose costs on others in the team.

Another application is to residential home owners in which 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 mutual monitoring 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 be 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 (Woodburn 1982, Knauft 1991, Boehm 1993, Bowles and Gintis 1998).

Given the apparently widespread nature of the problems of non-contractibility which it addresses and the welfare benefits it may make possible, it may be wondered why mutual monitoring is not ubiquitous in modern economies. 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 mutual monitoring. 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 mutual monitoring 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 (Galor and Zeira 1993, Loury 1981, Blanchflower and Oswald 1998, Black, de Meza and Jeffreys 1996), and the extent of cooperative forms of ownership of team assets (Legros and Newman 1996).[22] Because a given distribution of contracts and incentives may be constrained Pareto-optimal under some distributions of wealth but not under others, a particular distribution of wealth may preclude the evolution of allocationally superior systems of contract and incentives. Thus the assertion that the assignment of residual claimancy and control rights in market economies may be deduced from considerations of allocative efficiency is not generally valid.[23]

[22]We survey these cases in Bardhan, Bowles and Gintis (2000).

[23]This presumption is expressed in Alchian and Demsetz (1972), Kihlstrom and Laffont (1979), Grossman and Hart (1986), Holmström and Tirole (1988), and Hart and Moore (1990).

References

Alchian, Armen and Harold Demsetz, “Production, Information Costs, and Economic Organization,” *American Economic Review* 62 (December 1972):777–795.

Arnott, Richard, “Moral Hazard and Nonmarket Institutions,” *American Economic Review* 81,1 (March 1991):180–190.

Banerjee, Abhijit V., Timothy Besley, and Timothy W. Guinnane, “Thy Neighbor’s Keeper: The Design of a Credit Cooperative with Theory and a Test,” *Quarterly Journal of Economics* (May 1994):491–515.

Bardhan, Pranab, Samuel Bowles, and Herbert Gintis, “Wealth Inequality, Credit Constraints, and Economic Performance,” in Anthony Atkinson and François Bourguignon (eds.) *Handbook of Income Distribution* (Dortrecht: North-Holland, 2000).

Besley, Timothy and Stephen Coate, “Group Lending, Repayment Incentives and Social Collateral,” *Journal of Development Economics* 46 (1995):1–18.

Bingham, Paul M., “Human Uniqueness: A General Theory,” *Quarterly Review of Biology* 74,2 (June 1999):133–169.

Black, Jane, David de Meza, and David Jeffreys, “House Prices, the Supply of Collateral and the Enterprise Economy,” *Economic Journal* 106 (January 1996):60–75.

Blanchflower, David and Andrew Oswald, “What Makes a Young Entrepreneur?,” *Journal of Labor Economics* 16,1 (January 1998):26–60.

Blount, Sally, “When Social Outcomes Aren’t Fair: The Effect of Causal Attributions on Preferences,” *Organizational Behavior & Human Decision Processes* 63,2 (August 1995):131–144.

Boehm, Christopher, “Egalitarian Behavior and Reverse Dominance Hierarchy,” *Current Anthropology* 34,3 (June 1993):227–254.

Bowles, Samuel and Herbert Gintis, “The Evolution of Strong Reciprocity,” 1998. Santa Fe Institute Working Paper #98-08-073E.

Camerer, Colin and Richard Thaler, “Ultimatums, Dictators, and Manners,” *Journal of Economic Perspectives* 9,2 (1995):209–219.

Cameron, Lisa, “Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia,” 1995. Discussion Paper no. 345, Department of Economics, Princeton University.

Campbell, Donald T., “Two Distinct Routes beyond Kin Selection to Ultra-Sociality: Implications for the Humanities and Social Sciences,” in Diane L. Bridgeman (ed.) *The Nature of Prosocial Development* (New York: Academic Press, 1983) pp. 11–41.

Coleman, James S., “Free Riders and Zealots: The Role of Social Networks,” *Sociological Theory* 6 (Spring 1988):52–57.

Craig, Ben and John Pencavel, “Participation and Productivity: A Comparison of Worker Cooperatives and Conventional Firms in the Plywood Industry,” *Brookings Papers: Microeconomics (*1995):121–160.

Croson, Rachel, “Partners and Strangers Revisited,” *Economic Letters* 53 (1996):25–32.

Davis, Douglas D. and Charles A. Holt, *Experimental Economics* (Princeton, NJ: Princeton University Press, 1993).

Dong, Xioa-yuan and Gregory Dow, “Does Free Exit Reduce Shirking in Production Teams?,” *Journal of Comparative Economics* 17 (1993):472–484.

— and — , “Monitoring Costs in Chinese Agricultural Teams,” *Journal of Political Economy* 101,3 (1993):539–553.

Falk, Armin and Urs Fischbacher, “A Theory of Reciprocity,” 1998. Unpublished Manuscript, Institute for Empirical Economic Research, University of Zurich.

Fehr, Ernst and Klaus M. Schmidt, “A Theory of Fairness, Competition, and Cooperation,” *Quarterly Journal of Economics* 114 (August 1999):817–868.

— and Simon Gächter, “Cooperation and Punishment,” *American Economic Review* 90,4 (September 2000).

— , — , and Georg Kirchsteiger, “Reciprocity as a Contract Enforcement Device: Experimental Evidence,” *Econometrica* 65,4 (July 1997):833–860.

Forsythe, Robert, Joel Horowitz, N. E. Savin, and Martin Sefton, “Replicability, Fairness and Pay in Experiments with Simple Bargaining Games,” *Games and Economic Behavior* 6,3 (May 1994):347–369.

Frohlich, Norman, John Godard, Joe Oppenheimer, and Frederick Starke, “Employee vs. Conventionally Owned and controlled Firms: An Experimental Analysis,” *Managerial and Decision Economics* 19 (1998):311–326.

Galor, Oded and Joseph Zeira, “Income Distribution and Macroeconomics,” *Review of Economic Studies* 60,1 (1993):35–52.

Gintis, Herbert, “Strong Reciprocity and Human Sociality,” *Journal of Theoretical Biology* 206 (2000):169–179.

Glaeser, Edward L. and DiPasquale, “Incentives and Social Capital: Are Homeowners Better Citizens?,” *Journal of Urban Economics* 45,2 (1999):354–384.

Greenberg, Edward, *Workplace Democracy: The Political Effects of Participation* (Ithaca: Cornell University Press, 1986).

Grossman, Sanford J. and Oliver D. Hart, “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration,” *Journal of Political Economy* 94,4 (August 1986):691–719.

Güth, Werner and Peter Ockenfels, "Efficiency by Trust in Fairness? Multiperiod Ultimatum Bargaining Experiments with an Increasing Cake," *International Journal of Game Theory* 22,1 (1993):51–73.

Güth, Werner and Reinhard Tietz, "Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results," *Journal of Economic Psychology* 11 (1990):417–449.

Güth, Werner, R. Schmittberger, and B. Schwarz, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3 (May 1982):367–388.

Hansen, Daniel G., "Individual Responses to a Group Incentive," *Industrial and Labor Relations Review* 51,1 (October 1997):37–49.

Hart, Oliver and John Moore, "Property Rights and the Nature of the Firm," *Journal of Political Economy* 98,6 (December 1990):1119–1158.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith, "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology," *Economic Inquiry* 36,3 (July 1998):335–352.

Holmström, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics* 7 (1982):324–340.

— and Jean Tirole, "The Theory of the Firm," in R. Schmalensee and R. Willig (eds.) *Handbook of Industrial Organization* (Amsterdam: North-Holland, 1988).

Homans, George, *Social Behavior: Its Elementary Forms* (New York: Harcourt Brace, 1961).

Hossain, M., "Credit for Alleviation of Rural Poverty: the Grameen Bank in Bangladesh," 1988. International Food Policy Research Institute Report 65.

Kandel, Eugene and Edward P. Lazear, "Peer Pressure and Partnerships," *Journal of Political Economy* 100,4 (August 1992):801–817.

Kihlstrom, Richard E. and Jean-Jacques Laffont, "A General Equilibrium Entrepreneurial Theory of Firm Formation Based on Risk Aversion," *Journal of Political Economy* 87,4 (1979):719–748.

Knauft, Bruce, "Violence and Sociality in Human Evolution," *Current Anthropology* 32,4 (August–October 1991):391–428.

Krebs, Dennis L. and Kathy Denton, "Social Illusions and Self-Deception: The Evolution of Biases in Personal Perception," in Jeffrey A. Simpson and Douglas T. Kenrick (eds.) *Evolutionary Social Psychology* (Mahwah, New Jersey: Lawrence Erlbaum Associates, 1997) pp. 21–47.

Laffont, Jean Jacques and Mohamed Salah Matoussi, "Moral Hazard, Financial Constraints, and Share Cropping in El Oulja," *Review of Economic Studies* 62 (1995):381–399.

Ledyard, J. O., “Public Goods: A Survey of Experimental Research,” in J. H. Kagel and A. E. Roth (eds.) *The Handbook of Experimental Economics* (Princeton, NJ: Princeton University Press, 1995) pp. 111–194.

Legros, Patrick and Andrew F. Newman, “Wealth Effects, Distribution, and the Theory of Organization,” *Journal of Economic Theory (*August 1996).

— and Steven A. Matthews, “Efficient and Nearly-Efficient Partnerships,” *Review of Economic Studies* 68 (1993):599–611.

Loury, Glen, “Intergenerational Transfers and the Distribution of Earnings,” *Econometrica* 49 (1981):843–67.

McAfee, R. Preston and John McMillan, “Optimal Contracts for Teams,” *International Economic Review* 32,3 (August 1991):561–577.

Ostrom, Elinor, *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge, UK: Cambridge University Press, 1990).

— , James Walker, and Roy Gardner, “Covenants with and without a Sword: Self-Governance Is Possible,” *American Political Science Review* 86,2 (June 1992):404–417.

Rotemburg, Julio J., “Human Relations in the Workplace,” *Journal of Political Economy* 102,4 (1994):684–717.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, “Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study,” *American Economic Review* 81,5 (December 1991):1068–1095.

Sampson, Robert J., Stephen W. Raudenbush, and Felton Earls, “Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy,” *Science* 277 (August 15, 1997):918–924.

Stiglitz, Joseph E., “Peer Monitoring and Credit Markets,” in Karla Hoff, Avishay Braverman, and Joseph E. Stiglitz (eds.) *The Economics of Rural Organization: Theory, Practice, and Policy* (New York: Oxford University Press, 1993) pp. 70–85.

Tilly, Charles, “Charivaris, Repertoires and Urban Politics,” in John M. Merriman (ed.) *French Cities in the Nineteenth Century* (New York: Holmes and Meier, 1981) pp. 73–91.

Varian, Hal R., “Monitoring Agents with Other Agents,” *Journal of Institutional and Theoretical Economics* 46,1 (1990):153–174.

Verba, Sidney, Kay Lehman Schlozman, and Henry Brady, *Voice and Equality: Civic Voluntarism in American Politics* (Cambridge, MA: Harvard University Press, 1995).

Weissing, Franz and Elinor Ostrom, “Irrigation Institutions and the Games Irrigators Play: Rule Enforcement without Guards,” in Reinhard Selten (ed.) *Game Equilibrium Models II: Methods Morals and Markets* (Berlin: Springer-Verlag, 1991) pp. 188–262.

Whyte, William F., *Money and Motivation* (New York: Harper & Row, 1955).

Woodburn, James, “Egalitarian Societies,” *Man* 17,3 (1982):431–451.

7 Appendix - Participant Instructions

You have been asked to participate in an economics 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 Francs instead of Dollars. Your payoffs will be calculated in terms of Francs and then translated at the end of the experiment into dollars at the following rate: 30 Francs = 1 Dollar.

Each participant receives a lump sum payment of 15 Francs at the beginning of the experiment (on top of the $5.00 show-up payment). This one-time payment may be used to offset any losses that are incurred during the experiment. However, it should be noted that you can ALWAYS avoid losses through your own decisions.

The experiment is divided into 10 different periods. In each period participants are divided into groups of 5. You will therefore be in a group with 4 other participants. The composition of the groups will change randomly at the beginning of each period. Therefore, in each period your group will consist of different participants.

Each period of the experiment consists of two stages. In the first stage you will decide how many francs you want to invest in each of two investment accounts. One account is a Private Account, which only you benefit from. The second account is a Public Account, the benefits of which are shared equally by all members of your group. In the second stage of the period you will be shown the investment behavior of the other members of your group. You can then decide whether you want to distribute points to members of your group. If you distribute points to other members of your group, their earnings will be reduced.

Now we will explain the two stages in more depth.

Stage One

At the beginning of every period each participant receives and endowment of 20 francs. You have to decide how much of this endowment you want to invest in each of the two accounts mentioned above. You are asked to invest in whole franc amounts (i.e. an investment of 5 francs is alright, but 3.75 should be rounded up to 4).

To record your investment decision, you will type the amount of francs you want to invest in the Public and/or the Private account by typing in the appropriate text-input box which will be yellow. Once you have made your decision, there will be a green Submit button that will record your investment 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. Your return on your Private Account. Your Private Account returns 1 franc for each franc invested. That is, for each franc invested in the Private Account you get 1 franc back.

2. Your return from the Public Account. Your earnings (and everyone else's in your group) is equal to .3 times the total investment by all members of the group to the Public Account.

Your Earnings can be summarized as follows:

1 ×(Investment in Private Account) + .3×(Group Total Investment in Public Account)

The income of each group member from the Public Account is calculated the same way. This means that each group member receives the same amount from the total investment in the Public Account. For example, consider the case of groups with 5 members, if the total investment in the Public Account is 75 francs (e.g. first group member invests 15 francs, the second 20, the third 10 and the fourth and fifth 15 each) then each group member will receive $.3 \times 75 = 22.5$ francs. If the total investment was 30 francs then each group member would receive $.3 \times 30 = 9$ francs.

For each franc you invest in the Private Account you get 1 franc back. Suppose however you invested this franc in the Public Account instead. Your income from the Public Account would increase by $.3 \times 1 = .3$ francs. At the same time the earnings of the other members of your group would also increase by .3 francs, so the total increase in the group's earnings would be 1.5 francs. Your investment in the Public Account therefore increases the earnings of the other group members. On the other hand your earnings increase for every franc that the other members of your group invest in the Public Account. For each franc invested by another group member you earn $.3 \times 1 = .3 francs$.

Stage Two

In stage two you will be shown the investment decisions made by other members of your group and they will see your decision. Also at this stage you can reduce the earnings made by other member of your group, if you want to. You will be shown how much EACH member of your group invested in both the Public and Private Accounts. Your investment decision will also appear on the screen and will be labeled as 'YOU'. Please remember that the composition of your group will change at the beginning of each period and therefore you will not be looking at the same people all the time.

You must now decide how many points (if any) you wish to give to each of the other member of your group. You distribute points by typing them into the input-text box that appears below the investment decision of each of the other group members.

You will have to pay a cost to distribute points to other group members. This cost increases as you distribute more points to another participant. You can distribute between 0 and 10 points to each other member of your group. Your total cost of distributing points is the sum of all the costs you incur for distributing points to each of the other group members. The following table illustrates the relationship between the points distributed to each group member and the costs of doing so in francs.

```
Points:                      0 1 2 3 4 5  6  7  8  9 10
Cost of Points (in francs): 0 1 2 4 6 9 12 16 20 25 30
```

Consider the case where there are 5 people per group. Suppose you assign 2 points to a group member. This costs you 2 francs. If you assign 9 points to another group member, it will cost you 25 francs and if you assign 0 points to the rest of the members of your group, you do not incur any cost. In this case your Total Cost of assigning points is $(2 + 25 + 0 + 0)$ or 27 francs. At any time you will be able to calculate your total cost of distributing points by clicking the orange Calculate Cost button that will appear on the screen. When you have finished distributing points you will click the blue Done button.

If you assign 0 points to a particular group member you do not change his or her earnings. However, for each point you assign to a group member, you reduce his or her Gross Earnings in the current period by 10 percent. Hence, if you assign one group member 2 points, his or her Gross Earnings for the period will be reduced by 20%. Assigning 4 points reduces Gross Earnings by 40% etc.

How much a participant's earnings from the first stage are reduced is determined by the Total amount of points he or she receives from all the other group members. If a participant receives a total of 3 points (from all the other group members in the current period) then his or her Gross Earnings would be reduced by 30 percent. If someone is assigned 4 points in total his or her Gross Earnings would be reduced by 40 percent. If anybody is assigned 10 or more points their Gross Earnings will be reduced by 100 percent. In this case the Gross Earnings of this person would be 0 francs for the current period.

For example, if a participant had Gross Earnings of 30 francs from the first stage and was assigned 3 points in the second stage, then his or her earnings would be reduced to $30 - (.3 \times 30) = 30 - 9 = 21$ francs.

In general, your earnings after the second stage will be calculated as follows:

Total Earnings at the end of the Second Stage:

1. If you received fewer than 10 points then Total Earnings equal

(Gross Earnings from Stage One)-[Gross Earnings?(.1$\times$ received points)]$-$(the cost of points you distributed)

2. If you receive 10 or more points then Total Earnings equal − (the cost of the points you distributed)

Please note that your earnings at the end of the second stage can be negative, if the cost of the points you distribute exceeds your (possibly reduced) earnings from stage one. However, you can avoid such losses by the decisions you make.

After all participants have made their decisions in the second stage, your final earnings for the period will be displayed in a manner similar to what follows:

Earnings Screen at the end of the Period

Your Gross Profits in the Current Period: The Total Cost of the Points You Assigned to Others: Number of Points Assigned to You by Others: Reduction of Gross Profit due to Points Assigned to You: % Current Period Payoff after Subtractions: Your Accumulated Earnings Including this Period:

When you have finished reviewing your earnings for the current period you will click the orange Proceed to Next Period button and wait for others to finish. When everyone is done, the experiment will proceed to the next period starting with stage one.

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.

e\Papers\Mutual Monitoring\Mutual Monitoring.tex January 24, 2001