The Ecology of Collective Action: A Public Goods and Sanctions Experiment with Controlled Group Formation[*]

by Umut Ones and Louis Putterman[**]

Abstract

Mounting evidence suggests that the outcomes of laboratory public goods games, and collective action in firms, communities, and polities, reflect the presence in most groups of individuals having differing preferences and beliefs. We designed a public goods experiment with targeted punishment opportunities to (a) confirm subject heterogeneity, (b) test the stability of subjects' types and (c) test the proposition that differences in group outcomes can be predicted with knowledge of the types of individuals who compose those groups. We demonstrate that differences in the inclination to cooperate have considerable persistence, that differences in levels of cooperation after many periods of repeated interaction can be significantly predicted by differences in inclination to cooperate which are manifested in the initial periods, and that significantly greater social efficiency can be achieved by grouping less cooperative subjects with those inclined to punish free riding while excluding those prone to perverse retaliation against cooperators.

JEL #s: D91, D92, H41, D23
Keywords: public goods, voluntary contribution mechanism, heterogeneous preferences, group formation

Corresponding author: Louis Putterman, Department of Economics, Brown University, Providence, RI 02912. Louis_Putterman@Brown.Edu.

[*] We are indebted to Toby Page, who collaborated with Putterman on a number of related experiments, for his help in designing the experiments on which this paper reports. We thank Dennis Zachary Shubert for his work on the program with which the experiments were run. The research was funded by National Science Foundation grant SES-0001769.

[**] The authors are respectively candidate for the Ph.D. and professor of economics at Brown University.

The Ecology of Collective Action: A Public Goods Experiment with Controlled Group Formation

by Umut Ones and Louis Putterman

0. Introduction

The prisoners' dilemma game models a dynamic that is common to many economic interactions. Workers in teams, partners in firms, and communities of individuals who share a common environment or common interest often confront the dilemma that all cooperating toward a shared goal is in their joint interest, yet each is better off if others cooperate while she herself shirks responsibility. In some work teams, partnerships, irrigation associations, village woodlot projects, and other groups, collective action succeeds far beyond the expectations of the conventional free rider analysis; but in others, failure is the norm.

To better understand why collective action sometimes succeeds and at other times fails, economists have conducted dozens of experiments with an *n*-person linear public goods game known as the voluntary contribution mechanism (VCM). These studies exhibit a high degree of concurrence in finding that (a) in one shot public goods games and in the first period of repeated games, subjects contribute an average of 50% or more of their endowments to the public good, and (b) in repeated play, contributions tend to decline with repetition, reaching an average of 10 or 15% in an announced last period (for reviews of the literature, see Davis and Holt, 1993; Ledyard, 1995).

The natural question for economists to ask about these results was whether they made necessary a reconsideration of conventional game theory, or whether with a little effort they could be reconciled with it. The dominant strategy of a rational agent intending to maximize his or her own payoff only, faced with other players of the same type who have common knowledge of their types, is to contribute nothing to the public good. Since subjects in most VCM experiments can't contribute negative amounts, errors due to unfamiliarity with the game would produce a natural upward bias. The decay in contributions might thus be interpreted as evidence of learning. But experimentalists are coming to reject a pure learning interpretation in the face of evidence against it. Contributions regularly rise again, even for experienced subjects, when the game is restarted (Andreoni, 1988). When high contributors are grouped by the experimenter with other high contributors, their contributions are sustained at high levels (Gunnthorsdottir *et al*., 2002). When subjects have an opportunity to impose costly monetary punishment on specific others in their group (as opposed to punishing indiscriminately by withholding contributions), high contributors tend to continue contributing while punishing free riders, who respond by raising their contributions (Fehr and Gächter, 2000a, hereafter FG). Rather than being due to some "typical" subject learning with experience that it's best to free ride, argue FG, the usual fall-off of contributions in the standard VCM might better be attributed to the interactions between subjects prone to free riding and others more inclined to cooperate conditional on others' doing likewise.

FG's approach is one of several which suggest that the outcomes of public goods experiments can't be understood without recognizing the presence of subjects having different preferences. In addition to the *actual* presence of subjects whose subjective payoffs don't coincide with the material payoffs of the game, the existence of *beliefs* that such subjects may be present, and that other subjects may also believe that such types are present, can explain observed behaviors in a Bayesian model along the lines of Kreps, Wilson, Milgrom and Roberts (1982).[1] Andreoni's (1995) analysis leads him to conclude that "on average about half of all cooperation comes from subjects who understand free-riding but choose to cooperate out of some form of kindness." Offerman, Sonnemans and Schram (1996) and Palfrey and Prisbrey (1997) find evidence of "warm glow" giving, in which the donor acts as if obtaining utility from contributing to the public good irrespective of the benefit received by others. FG suggest that conditionally cooperative subjects reciprocate the "kind" contributions of other cooperators and punish the "unkind" free riding of self-interested types. Ahn, Ostrom and Walker (2002) argue that most behavior in public goods experiments can be explained by subjects having varying degrees of inequality aversion, with some subjects simply being payoff maximizers. Fischbacher, Gächter and Fehr (2001) and Kurzban and Houser (2001) identify many subjects in their conditional and circular contribution games as cooperators and conditional cooperators.

If subjects differ in type and if the decay of contributions typical in experiments with randomly formed groups is attributable to the way that conditional cooperators respond to free riders in the absence of punishment or partner selection mechanisms, then the study of group behaviors becomes a study of an *ecology of interacting types*.[2] Nature may have given rise to heterogeneity among human individuals because an ongoing interplay of types proved evolutionarily stable,[3] but humans may be able to design institutions that give more or less beneficial results by manipulating the types of individuals comprising particular groups, and by exposing people to social environments that may help (along with possibly varying in-born predisposition) to determine type (in biological usage, phenotype). As noted, Gunnthorsdottir *et al*. (2002) show in a basic VCM experiment that cooperative players can achieve superior outcomes when grouped together by the experimenter without knowing that this is being done. In this paper, we carry their approach further by controlling group formation in a more complex collective action environment in which subjects have not one but two decision-variables under their control. In their much-emulated experiment,[4] Fehr and Gächter (2000a) introduced a

[1] For models using this approach to explicitly show the viability of cooperative behaviors, see Guttman (2000), (2003).

[2] As in Schelling's famous "ecology of micromotives" (1971), the emergent properties of the social system are distinct from the intentions and not immediately predictable from the actions of the individuals involved. Schelling's discussion did not, however, emphasize preference heterogeneity.

[3] For a survey of evolutionary models of preference formation, including ones in which both reciprocator and payoff maximizing behaviors exist in equilibrium, see Sethi and Somanathan (2003). Heterogeneity of human behavioral inclinations may be due to differences of culture and individual upbringing, as well as of genes. See Boyd and Richerson (1985), Durham (1991) and Ben-Ner and Putterman (1998).

[4] Replications include Carpenter and Matthews (2002), Sefton, Shupp and Walker (2002), Fehr and Gächter (2002), Masclet, Noussair, Tucker and Villeval (2002), and Bochet, Page and Putterman (forthcoming).

degree of freedom into the VCM when they permitted subjects to impose costly punishments on one another after learning their contributions to a public good. If the propensity to punish is not perfectly correlated with the propensity to contribute to the public good, then the public goods game with punishment stage, which mirrors aspects of collective action in real world groups, will display a somewhat more complex ecology of subject types than can be detected in a game over contributions only.

In interpreting their results, FG emphasized that punishment was mainly given to low contributors by high ones, allowing them to describe subjects in terms of just two types—reciprocators, who both contribute and punish free riders, and payoff maximizers, who contribute only when they anticipate punishment and who never incur the cost of punishing others. But Cinyabuguma, Page, and Putterman (in process, hereafter CPP) find that, both in the cooperation-with-punishment experiments reported in Bochet, Page and Putterman (forthcoming, hereafter BPP) and Page, Putterman and Unel (2003, hereafter PPU) and in the FG (2000) experiment data,[5] about 20% of punishment is aimed at *high* contributors. They find that substantial numbers of *low* contributors imposed costly punishment, sometimes on other low contributors, sometimes on high contributors, and sometimes on both. Also, propensities to punish and propensities to contribute are imperfectly correlated: there are high contributors who never punish, and low contributors who frequently do so.[6] Anderson and Putterman (2004) also report similar results for a set of perfect stranger treatment cooperation with punishment experiments.

We designed an experiment to study the persistence of subject types, the relative influences of predisposition versus experience, and our ability to predict group outcomes by manipulating group composition. In our experiment, we find that the inclination to cooperate has considerable persistence, that differences in levels of cooperation after fifteen rounds of repeated interaction can be significantly predicted by differences in inclination to cooperate manifested in the initial rounds, and that significantly greater social efficiency can be achieved by grouping less cooperative subjects with those inclined to punish free riding while excluding those prone to perverse retaliation against cooperators.

Our paper proceeds as follows. In section 1, we discuss the theoretical framework of our study and its relationship to the existing literature. In section 2, we explain the design of our experiment. In section 3, we describe the results as they illustrate the general character of cooperation and punishment behaviors. Section 4 focuses on results with respect to differentiation among groups. Section 5 analyzes the persistence of individual types, evidence of environmental influences, and the relationship between

[5] Kindly provided by those authors.

[6] Some high contributors may not be active punishers because the propensities toward positive and toward negative reciprocity may have different strengths in different individuals, as discussed further, below. Punishment of low contributors by other low contributors is simply a matter of self-interest in a repeated partner-group setting, since the punisher's earnings rise if punishment induces the targeted individual to contribute more. That not all low contributors engage equally in punishment may reflect differences in their degree of strategic sophistication and different prior beliefs about the impact of punishment on others' contributions.

contributing and punishing propensity. Section 6 concludes with a discussion and summary.

1. Theoretical framework

a. Agent types

Several kinds of preferences, among them inequality aversion, altruism, and "warm glow," offer potential explanations for the persistence of positive contributions in public goods experiments. We focus on the preference called *reciprocity* or *conditional cooperation* because it is consistent with the actions of many subjects in recent voluntary contribution experiments, and because of the existence of a growing literature on the topic by anthropologists, sociologists, sociobiologists, evolutionary psychologists, and economic theorists and experimentalists.[7] According to Hoffman, McCabe and Smith (1998) and Fehr and Gächter (2000b), reciprocity entails an inclination to confer benefits on those who help one and to impose costs on those who harm one. The first, favor-returning part, can be called "positive reciprocity," the second, harm-returning part, "negative reciprocity." In both parts, the reciprocator shows a willingness to incur costs, and accordingly his or her actions are inconsistent with an exclusive preference for maximum material payoff. Although in repeated play with the same partner or when reputation carries into play with others, an agent may obtain material benefit in the long run if she creates an expectation of like future actions, true reciprocators reciprocate even in one shot games or end game situations with no potential for reputational gains.[8] Our experiment allows for the possibility that the strengths of the two faces of reciprocity are not perfectly correlated across individuals.

In a public goods game, reciprocity can affect behavior because the contribution of another person benefits one, which calls for conferring a benefit in return. Specifically, the material payoff of a group member i is given by

$$y_i = (E_i - C_i) + (m)\Sigma_{\text{all } j}\, Cj \qquad (1)$$

where E_i is i's endowment, C_i is i's contribution to the public good, the summation is taken over all members of the group, and m is a constant greater than $1/n$ and less than 1, with n being the number of members in the group. While this generates the material payoffs of a classic prisoners' dilemma, as in Figure 1a, with $c > a > d > b$, the psychic payoffs of the game are for the reciprocator what Sen (1967) called assurance game payoffs—i.e., $a > f > d > e$.

[7] Examples include Boyd and Richerson, 2002, Henrich and Boyd, 2001, Cosmides and Tooby, 1989, Rabin, 1993, Sethi and Somanathan, 2003, Gintis 2000, Guttman, 2000, Guttman, 2003, McCabe, Rassenti and Smith, 1996, Fehr and Gächter, 2002b, and Ben-Ner and Putterman (2002).

[8] Examples include engaging in punishment of low contributors in the perfect stranger conditions of Fehr and Gächter (2000a) and Anderson and Putterman (2003). Gintis, Bowles, Boyd and Fehr (forthcoming) call this propensity "strong reciprocity" to distinguish it from the kind of reciprocity motivated by self-interest.

Some expositions suggest a simple dichotomy of agents into reciprocators and non-reciprocators, the latter being strictly self-interested types. To understand the apparently limitless variety of behaviors actually observed, however, it's helpful to think in terms of *degrees* of reciprocity lying along a continuum. An agent's degree of reciprocity can be measured by the size of the private gain she is willing to forego in order to bestow a benefit on her benefactor—in terms of the figure, $c - f = \varphi$, or the psychic disutility of free-riding on others when the latter are cooperative. For a payoff maximizer, $f = c$ ($\varphi = 0$), and since $c > a$, the payoff maximizer will free ride if he knows his fellow player is contributing. The strongest reciprocators have the largest positive values of φ, leading to $c > a > f$, in which case it is subjectively better not to free ride when other group members aren't doing so; but other agents may have intermediate positive values of φ, so that f may be greater than or less than a. φ thus constitutes an index of reciprocity.[9]

	Cont.	Free-ride
Cont.	(a,a)	(b,c)
Free-ride	(c,b)	(d,d)

material pay-offs

	Cont.	Free-ride
Cont.	(a,a)	(e,f)
Free-ride	(f,e)	(d,d)

psychic pay-offs

Figure 1. *The contribution game has material pay-offs such that $c > a > d > b$ for all players. Pay-off maximizers are agents for whom psychic and material payoffs are identical. Strong reciprocators have psychic pay-offs $a > f > d > e$.*

One could logically suppose that reciprocity would not impact contributions to a public good unless the same group interacted repeatedly and could thus respond to one another's previous moves. Yet even when moves are simultaneous, one can alternatively assume that a reciprocator prefers to contribute if he believes that others are also doing so. On this interpretation, a reciprocator acts, even in a one shot game, according to his expectation or belief about what others will do. If reciprocators are optimistic about others' contributions at the outset and, in addition, trust that fellow players will continue to act as they have acted thus far, then they will behave in the VCM like tit-for-tat players, contributing on the first round and continuing to contribute in each subsequent one provided that others have done so thus far. It also follows that two individuals with the same degree of reciprocity might act differently owing to different beliefs. Indeed, they may differ not only in their beliefs about first period behaviors, but also in whether they tend to treat contributions by others as strong indications that the latter are reciprocators or rather to accord weight to the possibility that those others are

[9] Ahn, Ostrom and Walker (2002) find support for this interpretation by studying willingness to contribute despite the foregone private benefit through varying the size of that benefit as a treatment parameter.

opportunists feigning reciprocity and planning to defect at an advantageous time. To keep matters simple, we will denote the optimism or pessimism of beliefs by a simple scalar index, ***B***.

The task of predicting agents' behaviors when contributions can take more than one value and when groups contain more than two agents, as in our experiments, is clearly not an easy one. To predict precisely, we would not only need to know agents' initial beliefs and their rules for updating them, but also whether they would continue to contribute their entire endowment only if all other group members do so, whether it is sufficient that half of the others contribute fully, or that all others contribute at least half of their endowments, and so on. To cut through these complexities, we will *usually* interpret our findings with the aid of the simplifying assumption that beliefs and their updating differ mildly enough among subjects so that differences in contributions to the public good, for a given profile of contributions by other group members, are evidence of differing degrees of reciprocity. In other words, if subjects *i* contributes more than subject *j*, especially when playing in the same group with the same history, we will provisionally assume that *i* has more reciprocal preferences (higher φ) than *j*, although our discussion leaves open the possibility that *i* and *j* have the same φ (or even that *j* has higher φ than *i*) but that *i* has more optimistic beliefs about what others will contribute.

An important reason for treating differences in contributions as reflecting differences in degrees of reciprocity as provisional is that it may be in individuals' interests to *feign* reciprocity during some portion of finitely repeated play. Following the logic of Kreps *et al*., if a payoff-maximizer (non-reciprocator) believes that those with whom he interacts have an *ex ante* belief that the fraction of reciprocators in their group is above a critical value, self-interest dictates acting *like* a reciprocator so as to engender a series of cooperative moves, ending with a late-game defection.[10] Beliefs, including beliefs about others' beliefs, are thus important to the choices of payoff-maximizing players. Two individuals both of whom care only about own payoff may choose to contribute different amounts because they differ in their beliefs as to the proportion of reciprocators others believe to be present. For simplicity, we subsume both the belief about the proportion of actual reciprocators and beliefs about others' beliefs about that proportion in the same belief measure, ***B***.[11]

Negative reciprocity also has a part to play. Not contributing when others contribute may be seen as exploiting others' kindness, and this can trigger negative reciprocity—that is, a desire to punish the free rider—in the high contributor. In the basic VCM without punishment stage, negative reciprocity can only take the form of reducing one's own contributions, a blunt instrument since there is no way to direct it differentially againt free riders without also hurting high contributors in the group (Fehr

[10] Time preference differences could also play a role in events played out over time, but they can be disregarded when considering multiple rounds of play in the course of a brief experiment.

[11] How much a pay-off maximizer contributes to a public good might depend not only on his beliefs about others' types and beliefs, but also on his degree of strategic sophistication. We could denote the degree of strategic-mindedness or sophistication of subjects by another scalar index. But we will suppress this consideration to avoid excessive complexity.

and Gächter (2000a)). Moreover, because the signal can't be distinguished from ordinary free riding, it may provoke further contribution declines by others. Attributing the decay of contributions in the typical VCM experiment to this factor, Fehr and Gächter (2000a) designed an experiment in which a targeted punishment opportunity follows the contribution stage, thus allowing reciprocators to continue to engage in positive reciprocity toward high contributors—by contributing to the public good—while simultaneously engaging in negative reciprocity toward free riders—by imposing costly punishment on them. The result was that contributions rose rather than fell with repetition.

Although positive and negative reciprocity have been presented as two sides of the same coin, we see as an open matter, to be investigated empirically, how closely the strengths of the two tendencies are correlated. Some individuals may have a strong inclination to contribute to the public good if others also do so, yet they may have little or no inclination to punish free riders, whether because they are reluctant to hurt others, see this as a waste of resources, or are emotionally slow to anger. Other individuals may be more cautious about contributing to the public good, being willing to do so only if they see high contributions by all or most others, yet they may be quick to punish free-riders due to a stronger inclination to anger, a lower threshold for considering themselves to be exploited, less reservation about hurting others, etc. Rather than supposing the degree of reciprocity being captured by a single variable, then, we allow the degree of positive and that of negative reciprocity to be somewhat independent of one another. We denote the strength of the inclination to punish free riders by $\boldsymbol{N}$ (for negative reciprocity).

The presence of targeted punishment opportunities changes a simple public goods game into a two stage game with a different strategy set and the possibility of different outcomes. If all players are payoff maximizers and have common knowledge of this, then no one will incur a cost to punish free riding, so not contributing to the public good would remain the unique sub-game perfect equilibrium behavior in the contribution stage. However, a sufficiently strong expectation that some players have a propensity to direct large punishments at free riders (i.e., have high $\boldsymbol{N}$) may induce payoff maximizing players to contribute to the public good. FG found that in the same subject groups in which contributions declined monotonically with repetition in a simple VCM condition contributions actually rose with repetition in a condition in which a punishment stage is added. Similar results have been confirmed by other researchers (see above).

Just as the incentive to feign a preference can influence contributions, however, so also will that incentive affect punishing: some agents who actually care only for their own monetary payoff (in particular, who have $\boldsymbol{N} = 0$) might punish like a high-$\boldsymbol{N}$ subject for early rounds of a finitely-repeated VCM with continuing group membership. If, à la Kreps *et al*., the self-interested agent believes that others are prepared to believe that negative reciprocators are present in the group, then it can be profitable even for a low contributing payoff maximizer to punish other low contributors, since this may well lead the latter to raise their contributions, benefiting the punisher.[12] In the last period, of

[12] Note that in the experiments, a subject does not know *which* other subject is punishing him, so even if the low contributor receiving the punishment would have interpreted differently punishment by a high

course, a strategic punisher will be distinguishable from a subject with strong preference ***N***, because the former will never incur costs to punish.

Analysis of past VCM-with-punishment experiments indicate that one other preference must also be accounted for to understand the ecology of interacting types. CPP have demonstrated that about 20% of punishment in VCM-with-punishment experiments is aimed at high, rather than low, contributors, and most of this seems explicable either as attempts to retaliate against the punishment the agent has herself received, or as attempts to raise the punisher's *relative* earnings at the cost of his *absolute* earnings, a motivation called "spite" by Saijo and Nakamura (1995). In a repeated game with fixed group composition, retaliatory punishment could be a self-interested response intended to make it safe to continue free riding with less likelihood of being punished again. But some retaliatory and spiteful punishment appear to stem from a preference type in its own right. Perverse punishment is observed even in the last period of play, and in perfect stranger designs (see CPP's analysis of FG, and Anderson and Putterman, 2003). If we think of the "pro-social" preferences of positive and negative reciprocity as lying at, say, the right end of a continuum that includes strictly payoff maximizing (self-interested) preferences somewhere to its left, then the tendency to punish cooperators might usefully be thought of as an actively "anti-social" preference lying to self-interest's other side (like reciprocity, it is also not payoff increasing, but unlike reciprocity, it reduces others' payoffs in addition to one's own). We assume that the strength of the inclination to punish perversely can be measured by another scalar, ***S***.

In sum, we design and analyze our experiment on the theoretical assumptions that (a) agents are characterized by a continuum of reciprocity types, φ, ranging from payoff maximizing ($\varphi = 0$) to strongly reciprocating (large positive φ); that (b) agents differ also in their beliefs about the proportion of reciprocators in their group and about the beliefs others hold in this regard (***B***); that (c) propensities to reciprocate positively by contributing when others contribute and to reciprocate negatively by punishing when others free ride may be imperfectly correlated, with positive reciprocity the stronger tendency for some, negative reciprocity (***N***) the stronger one for others; and that (d) a small number of agents may have a preference for perverse and/or spiteful punishment (***S***) that will lead them to punish high contributors even in a one-shot situation or in the last period of a finitely repeated game.

In principle, we should be able to predict an individual i's contribution C_i to a public good (or at least their first contribution, before other players' actions have had a chance to affect theirs) as

$$C_i = f(\varphi_i, \boldsymbol{B}_i, \boldsymbol{N}_{-i}, \boldsymbol{S}_{-i}) \qquad (2)$$

where the subscripts on the last two terms indicate that they refer to the expected characteristics of individuals *other than i*. The amount of costly punishment i gives to some $j \neq i$ can in principle be predicted by

---

contributor than punishment by another low one, he lacks that information. The strategically punishing low contributor can therefore expect to have the same efficacy as any other punisher.

$$p_{ij} = g(C_j, C_{-j}, P, \varphi_i, \boldsymbol{B}_i, \boldsymbol{N}_i, \boldsymbol{S}_i) \quad (3)$$

where P stands for the price or cost of punishment to the punisher,[13] and this time all subscripts are those of *i*, except for C_j, the contribution of the prospective target of punishment, and C_{-j}, the contributions of those other than the target, which may come into play because *j*'s deservingness of punishment may be interpreted in the context of others' contribution choices as well as those of *i* and *j*'s own.[14]

b. The ecology of types

If *individuals* can be characterized in terms of their beliefs and preferences, it should be possible to predict the outcomes of *group* interactions in a VCM-with-punishment game by knowing what types of individuals constitute the group. This is straightforward in a few simple cases: for example, if a group is composed entirely of reciprocators who begin with optimistic beliefs about one another's types, they will establish from the outset and maintain even without communication an equilibrium of high contributions, since each will begin with a high contribution and will continue to contribute having had her expectations validated. For some heterogeneous groups, too, outcomes may be fairly predictable. Consider a group of four with the following composition: two agents are payoff maximizers; two agents are strong reciprocators, one inclined more strongly towards positive than negative reciprocity, the other inclined more strongly towards negative than positive reciprocity. Suppose all four agents believe there to be a 50% chance of encountering a reciprocator, and that each believes the others to have the same beliefs as herself. Though our prediction might lack precision, we could be fairly confident that this group would begin with a range of positive although probably not maximal contributions, and that it would exhibit relatively high contributions after a few periods of play, since the high-***N*** subject would from the outset punish any low contributions, and the payoff maximizers will adjust their contributions upwards to avoid punishment.

Now change the composition of this hypothetical group so that instead of two payoff maximizers there is one payoff maximizer and one agent inclined towards low contributions and perverse punishment (the other two members are as before). The group outcome is now more difficult to predict, since the negative reciprocator will tend to punish the perverse punisher, who will punish back in return. Since the retaliator resists the pressure to contribute more, so might the payoff maximizer. The positive reciprocator too will refrain from making maximal contributions, since some others are failing to do so. A good deal of punishment might end up being wasted without inducing a rise in contributions. Without being more precise, we can certainly expect this group to exhibit lower average contributions and earnings than the counterpart group that differs in the type of one member.

[13] Carpenter (2003) and Anderson and Putterman (2003) find that the amount of punishment purchased by the punisher is a decreasing function of its cost to her.

[14] Predictive power of both equations may be enhanced by allowing for the influence of past history within a group.

c. Types and environments

These examples illustrate that it is not only individuals' preferences and beliefs, but also the way in which these interact with the preferences and beliefs of the others with whom they are grouped, that will influence their individual choices and their groups' outcomes in continuing interactions. We find helpful a notion of type/environment interactions that parallels (without precisely corresponding to) the heredity/environment and genotype/phenotype notions common in the behavioral sciences and biology. An individual presumably enters an encounter, in life or in the laboratory, as a bearer of certain preferences and beliefs that constitute her type, parallel to the genotype in biology. The individual's behavior will remain the same or change over the course of the interaction depending on the actions of the others she encounters, her environment. Conceivably, an individual might have no persistent type to speak of, and might simply move from encounter to encounter trying out new behaviors and retaining those that work out only for as long as they do so. However, our working hypothesis is that there is some persistence of type, perhaps because of the important contributions of genes and early socialization. It's an empirical question just how much type persistence individuals display and how much their behaviors are altered by the environments they encounter.

2. Experimental design

In our experiment, subjects play a repeated linear public goods game with punishment stage in groups of four, with contribution and punishment decisions.[15] Every period, each subject *i* has 10 experimental dollars of which he contributes an integer amount, C_i, possibly 0, to a group account, and retains (10 - C_i), giving provisional earnings

$$y_{i,p} = (10 - C_i) + (0.4)\Sigma_{\text{all } j}\, Cj \qquad (4)$$

where the summation is taken over all four group members, *i* included. After individual contribution decisions are revealed, subjects can use current period earnings to reduce the earnings of other group members at a cost of 0.25 to the punisher for each experimental dollar lost by the person targeted. *i*'s final earnings for the period are thus

$$y_{i,f} = (10 - C_i) + (0.4)\Sigma_j\, Cj - (0.25)\Sigma_j\, R_{ij} - \Sigma_j\, R_{ji} \qquad (5)$$

[15] The set up is identical to the treatments in BPP and PPU with punishment and without communication or endogenous group formation, and is the same as in FG except that (a) in our experiment, the cost of punishment to the punisher is a constant fraction of its cost to the person punished, and punishment is imposed in dollar terms, rather than in percentages of pre-punishment earnings, and (b) our experiments do not also include a VCM-without-punishment condition which FG use for purposes of within-subject comparison. Although our fixed per unit cost of punishment differs from that in FG (2000), the same set-up is used in A fixed cost per monetary unit deducted is used in Fehr and Gächter (2002) and in Sefton, Shupp and Walker (2002).

where R_{ij} is the number of dollars by which *i* reduces *j*'s earnings, and conversely for R_{ji}. As in FG, other group members are identified to *i* by letters B, C and D, which switch randomly each period, to minimize vendettas.

In each of the twelve sessions constituting the present set of experiments, sixteen subjects, undergraduates at Brown University drawn from all disciplines, engaged in twenty five sets of contribution and punishment decisions. While they were unable to communicate with others and unable to tell which other individuals constituted their own group, each could see that fifteen other subjects were present, and the experimenter truthfully informed them that they would be put in one group for the first period, a possibly different group whose members would not change during periods 2 – 5, another fixed group for periods 6 – 15, and a final fixed group for periods 16-25, with some overlap of membership among these four groupings being possible but not certain. No information was provided about the basis of group formation. We analyze three treatments, which use different bases to group subjects.

Two treatments, dubbed the homogeneous-random (or HR) and the random-homogeneous (or RH) treatments, attempted first to identify subject "types" by allowing them to play the contribution and punishment game for five periods in groups of similarly diverse composition. When these five "diagnostic" periods ended, the subjects played in homogeneous groups made up of members of similar "type" to themselves for ten periods, and in randomly formed groups of no special composition for ten periods. In the HR treatment, play in homogeneous groups occurred in periods 6-15 and that in random groups in periods 16-25, while in the RH treatment, this order was reversed (see Table 1).

The five "diagnostic" periods worked as follows. Each subject first decided on an amount to contribute to the group account in period 1. The computer then showed the subject the contributions of the other three members of her group, each group having been made similarly diverse by being assigned as members one of the four highest first-round contributors in the session, one of the four lowest contributors, and so on, without the subjects' knowledge.[16] After the subject made her first-period decisions on punishment of her three counterparts, there was a second placement into groups again meant to be diverse, this time in terms *both* of contribution *and* of punishment behavior. The computer did this by calculating a *reduction index* which is a larger positive number the more the subject engaged in punishment of low contributors, a larger negative number the more she punished high contributors, and zero if she didn't punish at all (a more complete description is given in Appendix A). Subjects were given ranks from 1 to 16 both for their contribution level (with 1 representing the highest level, 16 the lowest) and their reduction index (1 representing the most punishment of free riders, 16 the most punishment of high contributors), the two rank numbers were added together, and the four subjects with the lowest combined ranks (who tend to be both high contributors and

[16] Since subjects are always shown others' contribution amounts only after all have submitted their decisions, and since the computer made the group assignment in a fraction of a second, a subject who had participated in one of the experiments of BPP or PPU, in which assignment to groups preceded contribution decisions, could not have detected any difference, that is that the assignments were made *after* first decisions in this experiment.

strong punishers of free riders) were assigned to *different* groups, the four with the next lowest combined ranks to different groups, and so on.[17] This assured, to the extent possible, that every group had both high and low contributors, both aggressive punishers of low contributors and non-punishers or perverse punishers.

At the end of period 5, the subjects are ranked for their average contribution over periods 1 to 5 as a whole, and for their average punishment index over periods 1 to 5 as a whole. These ranks for behavior in periods 1 to 5 were the basis of their group assignments during the homogeneous grouping portion of the session, whether that was periods 6-15 (as in HR) or periods 16-25 (as in RH). Thus, even in the RH treatment, only behaviors in periods 1 to 5 were used in identifying the "type" of the subject. In both the period 2 group assignment and the homogeneous (period 6 or 16) group assignment stages, the contribution and reduction ranks were added together. For homogeneous groupings, the four subjects with the lowest summed rank, and so on, were placed in the *same* group, whereas for heterogeneous grouping (for diagnostic play) the apparently like subjects were dispersed among different groups.

We placed subjects into groups that came as close as possible to being equally diverse, during the "diagnostic" periods, because only by doing so could we hope to distinguish between behaviors attributable to differences in group-mates and those attributable to differences with which a subject entered the experiment. In particular, with homogeneous or randomly formed groups, one subject might chose a lot of punishment for low contributors while a second did not do so, but this could be because the first subject saw both high and low contributions in his group whereas the second saw only high or only low ones. We chose not to regroup subjects between periods 2 and 5, even though such regrouping might have aided in maintaining of heterogeneity, because our goal was to study the kind of ongoing interactions represented by partner groups, during periods 6-15 and 16-25, and we thus thought it better to get our "diagnostic" reading on types within a partner-type environment.

It should be noted that with these desirable features of the diagnostic portion of our experiment comes a much steeper challenge to the persistence of subject type than occurs in the experiment of Gunnthorsdottir *et al*. In that experiment, subjects were grouped with like contributors from period 1 onwards, and were regrouped as often as necessary to preserve homogeneity, which provided each subject with immediate reinforcement of his initial inclination and gave no opportunity to observe contrary behavior. Therefore, if subjects persisted in making characteristically high or low contributions after five periods of playing with heterogeneous others, in our HR treatment, or after fifteen periods of playing with heterogeneous and then randomly

[17] If two or more subjects were tied in contributions or in punishment indices, they received the same rank, for example if two tied for second place, both were treated as having the rank 2.5. This way the ranks of al 16 subjects always summed to 16!, assuring that the contribution and punishment ranks had equal weight when added together. If there were ties with respect to the combined rank, for example between the fourth and the fifth ranked subject, the computer broke them randomly. The four lowest ranked, the next four lowest ranked subjects, and so on, were allocated among the four groups in a random order, so that it was *not* the case that one group always included the subjects with combined ranks 1, 5, 9, and 13 while another included those ranked 2, 6, 10 and 14, and so on.

assigned others, in our RH treatment, this is stronger evidence of persistence of heterogeneous types than the already impressive evidence in Gunnthorsdottir *et al*.

Our third treatment is a baseline against which to measure the effects of the exogenous grouping procedures in treatments HR and RH. In it, subjects were placed in possibly different groups in periods 1, 2, 6 and 16, always by a strictly random grouping process; hence, we use the acronym RR. As in the HR and RH treatments, group membership remained fixed during periods 2 – 5, 6 – 15, and 16 – 25. Exactly the same instructions were given and subjects' tasks were identical in all three treatments, so subjects were unaware of the basis of grouping, unaware that different treatments existed, unaware that they were participating in one treatment rather than another, and unaware when they had been placed in a high contributor or a low contributor group relative to other groups in their session (of which they could have no knowledge). We conducted four sessions of each treatment using a total of 192 subjects. The three treatments are summarized in Table 1. Instructions are provided in Appendix B.

Table 1. Summary of treatments and grouping procedures.

	Treatment			**Overall**
Period(s)	**HR**	**RH**	**RR**	
1	Heterogeneous Groups[a]	Heterogeneous Groups[a]	Random Groups	
2 – 5	Heterogeneous Groups[b]	Heterogeneous Groups[b]	Random Groups	
6 – 15	Homogeneous Groups[c]	Random Groups	Random Groups	
16 – 25	Random Groups	Homogeneous Groups[c]	Random Groups	
Number of Sessions	4	4	4	12
Number of Subjects	64	64	64	192

Notes: a. Grouping based on initial contribution rank. b. Grouping based on initial contribution and initial punishment rank. c. Grouping based on combined contribution and punishment ranks of periods 1 – 5.

3. Results, Part I: General description.

Figure 2 plots the average contribution of subjects in the HR sessions, by period, while Figure 3 does the same for the RH sessions. For purposes of comparison, the average contribution of subjects in the RR sessions is also shown in each figure. In the figures, contributions are averaged over all subjects in periods in which there is no basis for distinguishing the 16 groups involved, but for those ten periods in which grouping was intendedly homogeneous by observed type, a separate line is shown for the average contribution of highest, second, third, and lowest ranked groups (called "Group 1," etc.). These separate lines average the contributions in the respective groups of each rank over the four sessions.

Result 1. *Contributions began at high levels and increased with repetition.*

The first thing to be observed is that in all three treatments, contributions to the group account began at relatively high levels, about 65% in HR, 74% in RR and 80% in

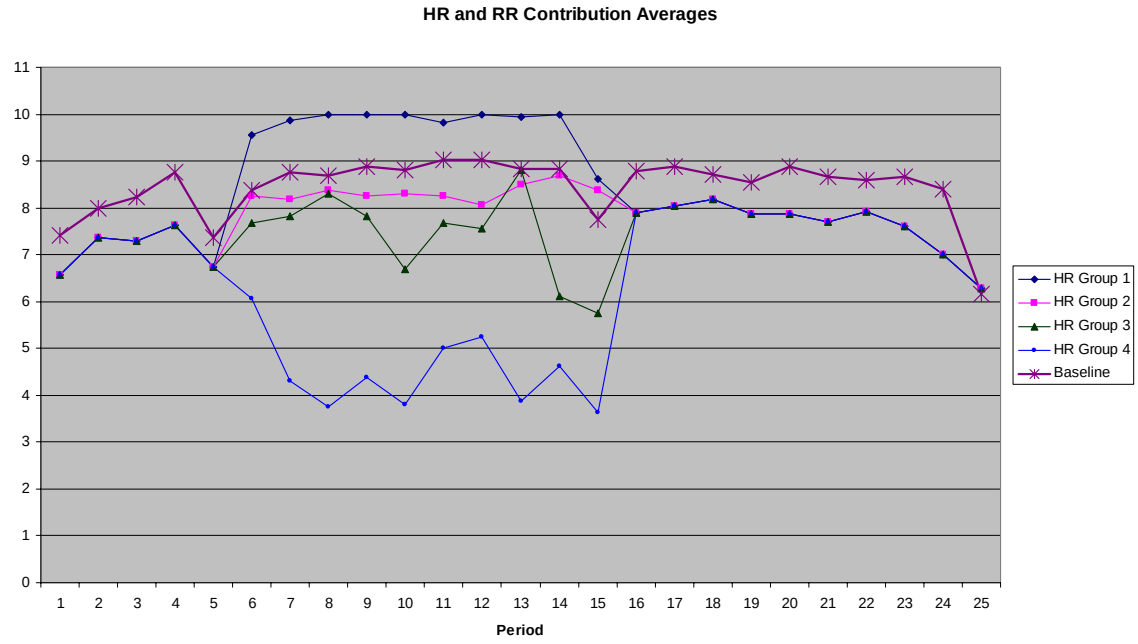

Figure 1. HR and RR (Baseline) average contribution by period and group.

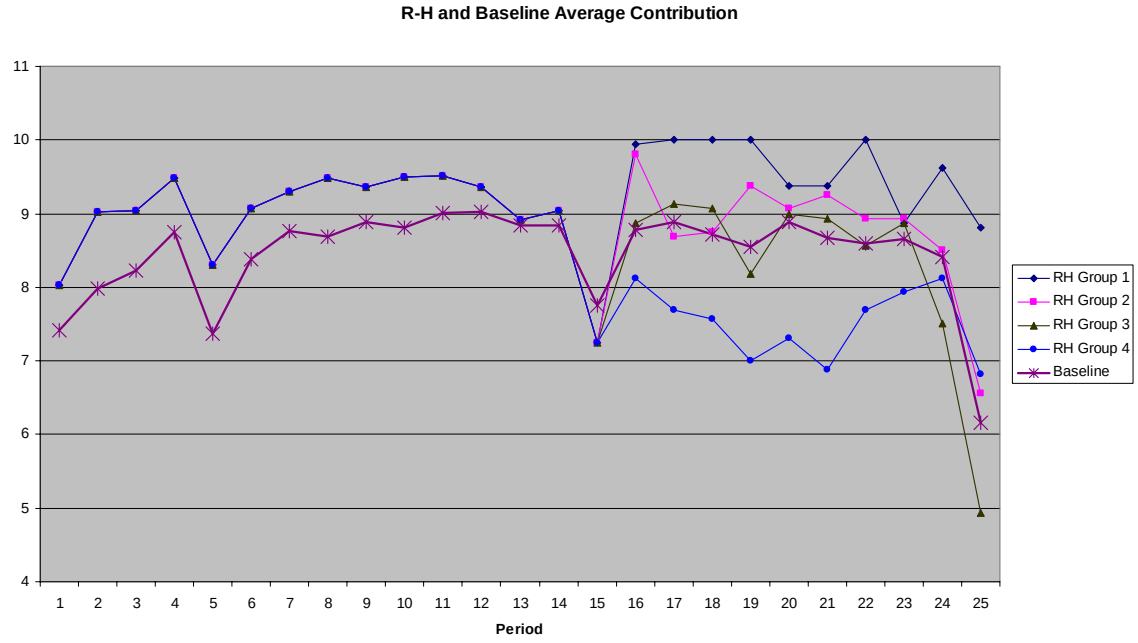

Figure 2. RH and RR (Baseline) average contribution by period and group.

RH, and that unlike in VCM experiments without punishment but like ones with punishment (FG, BPP, PPU, etc.), contributions do not show a tendency to decline with repetition. These two features suggest (a) that punishment of low contributions was anticipated even before it was observed by subjects,[18] and (b) that the threat of punishment, perhaps combined with reciprocators' abilities to continue contributing even while punishing free riders, kept most subjects contributing at relatively high levels, although small end-game effects are apparent in periods 5 and 15 and a larger one in the approach to period 25.[19] Discussion of the differences among groups that are observable in Figures 2 and 3 is postponed until Section 4.

Punishment was common in the experiment, including the last period, with 15%, 11%, and 11% of each subject's three opportunities to reduce others' earnings each period being utilized in the HR, RH, and RR treatments, respectively—equivalent to an average subject punishing some other subject in 1/3 or more of all periods. Most punishment was directed at low contributors, suggesting the presence of preference ***N***, but 35% of punishment dollars in the HR treatment, 38% in the RH treatment, and 10% in the RR treatment were perversely aimed at groups' highest contributors of the period, suggesting the presence of preference ***S*** as well.

Result 2. *As in other such experiments, the less an individual contributed to the group account relative to others, the more he or she tended to be punished. But in the HR and RR treatments, one was also more likely to be punished the more one exceeded the average, a clear indication of "perverse" punishment.*

Table 2 shows OLS estimates of a regression equation that follows a specification in FG, where the dependent variable is the number of dollars of reductions targeted at subject *j*, and the explanatory variables are (a) the *absolute negative deviation* of *j*'s contribution, defined as the difference between it and the average of other group members in the period, if *j* contributed less than the average, and as zero otherwise, (b) *j*'s *absolute positive deviation*, defined conversely, and (c) the others' average contribution.[20] For the HR and RR treatments, both deviation terms have highly significant coefficients, although the magnitude is higher for negative deviation, indicating that almost four times as much punishment was received per dollar below the average as per dollar above the average. Still, it is noteworthy that one would have been

[18] First period contributions in all three treatments are higher than in the similar no punishment treatments in PPU, as is true also for the treatments with punishment in the experiments of that paper. The average first period contribution is 6.0 in the 20 round no-punishment treatment of PPU, versus 6.6 in the HR treatment, 8.0 in the RH treatment, and 7.4 in the RR treatment. Mann-Whitney tests show the differences in first period contributions between our treatments and PPU's no punishment treatment to be insignificant for the HR treatment, but statistically significant at the 5% level for the RH treatment and at the 10% level for the RR treatment. We cannot account for the rather large difference in the first period contributions in the RH versus the HR treatment; these should have been about the same, since there were neither differences in instructions nor any differences that could have been induced by grouping algorithms before period 6.

[19] The rise of contributions in all three treatments in period 2 may be attributed mainly to some low contributors reacting to not-fully-anticipated punishment.

[20] All of the regressions are reported with robust (Huber-White) standard errors calculated using the robust command in Stata.

"sticking one's neck out" to contribute "too much" in some groups. The presence of this much perverse punishment is an obvious disincentive to efficiency and thus a reason why one might want to "exclude" perverse punishers from groups, as should happen in most groups during periods of homogeneous grouping. For the RH treatment, only the negative deviation term is significant (as in FG), perhaps in part because there was little margin above the average contribution most of the time. The point estimate of the coefficient on absolute positive deviation is nevertheless also positive.

	HR	**RH**	**RR**
Constant	0.429 ** (0.21)	0.844 * (0.52)	-0.979 (0.61)
Absolute Positive Derivation	0.086 *** (0.03)	0.032 (0.06)	0.137 ** (0.07)
Absolute Negative Derivation	0.534 *** (0.04)	0.572 *** (0.06)	0.910 *** (0.06)
Average Others' Contribution	0.004 (0.02)	-0.053 (0.05)	0.117 * (0.06)
Number of Observations	1600	1600	1600
$\mathbf{R^2}$	0.251	0.253	0.511
F Value	57.64 ***	35.40 ***	87.07 ***

Dependent Variable: Punishment Received
(an increase is a positive value)
(*) Significant in 10 %
(**) Significant in 5%
(***) Significant in 1 %
Cases where everyone in the group contributes the same amount are omitted. Numbers in parentheses are adjusted standard errors.

Table 2. Punishment received as a function of contribution deviations and average.

Result 3. *Subjects responded predictably to being punished.*

In the regressions in Table 3, one for each treatment, the dependent variable is the change in subject *j*'s contribution from period *t* to period *t*+1. The independent variables are formed by multiplying the amount of punishment received by *j* in period *t* by two dummy variables, the first being 1 if *j* contributed an amount equal to the group's average in the period or higher, 0 otherwise; the second being 1 if *j* contributed less than the group average, 0 otherwise. The coefficients can be interpreted as the impact of one dollar of punishment received, assuming *j*'s relative contribution was in the indicated category, upon *j*'s change in contribution. The estimated coefficients show that on average, a subject increased her contribution by somewhere between 51 and 68 cents for

every dollar of punishment she received if she had been contributing less than the group's average, consistent with the subject interpreting the punishment as an indication of displeasure for free riding or as a warning that more punishment would be forthcoming if the behavior continued. By contrast, subjects who were contributing the average or more when they received punishment tended to reduce their contributions by somewhere between 13 and 25 cents for each dollar of punishment received, which demonstrates the efficiency reducing impact of perverse punishment.

	HR	**RH**	**RR**
Constant	-0.436 *** (0.08)	-0.314 *** (0.06)	-0.504 *** (0.07)
Pun Received as High Contributor	-0.128 ** (0.06)	-0.245 *** (0.07)	-0.170 (0.16)
Pun Recieved as Low Contributor	0.617 *** (0.04)	0.682 *** (0.05)	0.513 *** (0.04)
Number of Observations	1536	1536	1536
$\mathbf{R^2}$	0.187	0.207	0.223
F Value	124.36 ***	110.78 ***	90.52 ***

Dependent Variable: Change in Contribution after the Punishment Received
(an increase is a positive value)
(*) Significant in 10 %
(**) Significant in 5%
Numbers in parentheses are adjusted standard errors.

Table 3. Effect of receiving punishment on the change in contribution

Result 4. *There is evidence of the negative reciprocity preference* ***N*** *in the form of last period punishing that cannot be explained by strategic motivations.*

A payoff maximizing subject would never punish in the last period. We can thus investigate whether free riders and high contributors were punished mainly to try to get them to change their behaviors, hence out of strategic motivation, or whether underlying preferences ***N*** and ***S*** were displayed by punishing in the last period of a group's interaction. We can confirm by a simple count that there were similar amounts of punishment in period 25 (and the last periods for specific groups, periods 1, 5, and 15) as in other periods. However, a more rigorous test can be done by estimating regression equations similar to those in Table 2 but including interaction terms to check whether the likelihood of being punished for a positive or negative deviations was any different in final periods. In one set of regressions, shown in Table 4, we multiply the absolute positive and absolute negative deviation by a dummy variable, DUMMY LAST, which

takes the value 1 in periods 1, 5, 15 and 25 and is 0 otherwise. In another set of regressions, we use one dummy variable, DUMMY 1,5,15, for periods that represent the last time a particular group plays together but not the last period in the session as a whole, and a separate dummy variable, DUMMY25, for the last period of the session, to test whether there is a stronger reduction of punishing in that period.

	HR		RH		RR	
Dependent Variable Punishment Received						
Constant	0.427 ** (0.21)	0.427 ** (0.21)	0.842 (0.52)	0.773 (0.55)	-1.045 (0.66)	-1.031 (0.65)
Absolute Positive Deviation	0.078 ** (0.04)	0.079 ** (0.04)	0.095 (0.07)	0.102 (0.07)	0.14 ** (0.07)	0.138 ** (0.07)
Absolute Negative Deviation	0.537 *** (0.05)	0.538 *** (0.04)	0.514 *** (0.06)	0.515 *** (0.06)	0.893 *** (0.06)	0.893 *** (0.06)
Others Contribution	0.004 (0.02)	0.004 (0.02)	-0.048 (0.05)	-0.041 (0.06)	0.124 ** (0.07)	0.123 ** (0.07)
Dummy All * Abs. Pos Dev	0.049 (0.06)		-0.145 *** (0.05)		0.012 (0.03)	
Dummy All * Abs. Neg Dev	-0.020 (0.09)		0.159 *** (0.13)		0.068 (0.16)	
Dummy 25* Abs. Pos Dev		-0.084 (0.08)		-0.205 ** (0.10)		0.071 (0.07)
Dummy 25*Abs. Neg Dev		0.032 (0.15)		0.362 (0.26)		0.136 (0.27)
Dummy 1,5,15*Abs. Pos. Dev.		0.101 (0.08)		-0.118 ** (0.06)		-0.021 (0.02)
Dummy 1,5,15*Abs. Neg. Dev.		-0.042 (0.11)		0.076 (0.13)		0.029 (0.17)
Number of Observations	1600	1600	1600	1600	1600	1600
$\mathbf{R^2}$	0.251	0.252	0.261	0.267	0.512	0.513
Dummy Significance (F)	4.256**	17.026***	34.511***	121.62***	6.532***	26.152***

(*)Significant in 10 %

(**)Significant in 5 %

(***)Significant in 1 %

Numbers in parentheses are adjusted standard errors.

Table 4. Punishment received as a function of contribution deviations.

Inspecting Table 4, we find no qualitative change with respect to the non-interacted deviation variables. Turning to the interaction terms, the DUMMY ALL interaction, covering all four last periods of partner groups, shows no significant difference except in the RH treatment, where it indicates that there was significantly *less*

perverse punishment and significantly *more* punishment of low contributors in last periods for that treatment. In the other specification, there are no significant interaction dummies except for the interactions of both DUMMY25 and DUMMY1,5,15 with absolute positive deviation. Consistent with the corresponding DUMMY ALL term for RH, these coefficients show there to be significantly *less* perverse punishment in last periods of that treatment. The interactions with absolute negative deviation are always insignificant, and for DUMMY25 they have positive values. These results support FG's belief that punishment of low contributors is not mainly strategic in nature, but is due to a preference (see also Falk, Fehr and Fischbacher (2001). The result for the interaction with positive deviation in the RH treatment raises some doubt as to whether the impulse to punish perversely is anything other than strategic, but since this effect is not observed in the other two treatments, our assumption that some subjects hold a preference ***S*** continues to have some support.[21, 22]

4. Results, Part II: Differences among groups and persistence of behaviors.

In this section, we investigate whether group behaviors were differentiated in our experiment as would be predicted if we correctly identified subject types and if types were persistent.

Result 6*: During the homogeneous grouping periods of both the HR and RH treatments, the ordering of contributions averaged over all sessions of each of the two treatments is exactly as would be expected assuming correct identification of types and persistence of the associated behaviors.*

---

[21] The negative but insignificant coefficient on DUMMY25*Abs. Pos. Dev. in the HR equation, and the quite similar magnitude of that coefficient to the coefficient on Absolute Positive Deviation itself, also hint at the idea that seemingly perverse punishment is really a strategic act to ward off future punishment for low contributions. However, the estimate for the RR treatment does not share this quality. The substantial perverse punishment exhibited in the perfect stranger experiment of Anderson and Putterman (2004) is also inconsistent with interpreting perverse punishment as mainly a strategy to increase payoff.

[22] Although figures 2 and 3 show that there were substantial contributions to the group account in period 25 in all three treatments, this can't be taken as definitive proof of *positive* reciprocity, because even payoff maximizing subjects had reason to contribute in the last period if they believed they would otherwise have been heavily punished. Last period contributions are consistent with, if not proof of, the possibility that many high contributors did have some degree of actual reciprocity ($\varphi > 0$). There were 5 subjects in the HR treatment who contributed their full endowment of 10 every period from period 2 to period 24; of these, 4 contributed 10 in period 25 as well. 11 subjects, including the 5 just mentioned, departed from contributing 10 no more than once during periods 2 to 24; of these, 8 continued to contribute 10 in period 25. More broadly, 18 HR subjects departed from contributing 8 or more no more than once in periods 2 to 24, and of these, 17 also contributed 8 or more in period 10. Similarly high proportions of high contributors maintained their high contributions in the last period of the RH treatment, which had more high contributors throughout. For example, 19 RH subjects departed from contributing 10 no more than once during periods 2 to 24, and of these, 15 contributed 10 in period 25 also. 35 RH subjects, including the 19, contributed 8 or more at least 22 times during periods 2 to 24; 28 of the 35, or 80%, also contributed 8 or more in period 25. More definitive evidence of positive reciprocity can be seen in the last period contributions to the public good in the endogenous group formation experiments of PPU, where in most groups subjects contributed an average of 50% or more of their endowments in the last period, compared to only 10% in the last period of a baseline treatment in which willingness to cooperate declined early in the face of inability to exclude free riders.

Figure 3

H-R Group Contribution Averages and RR Contribution Average

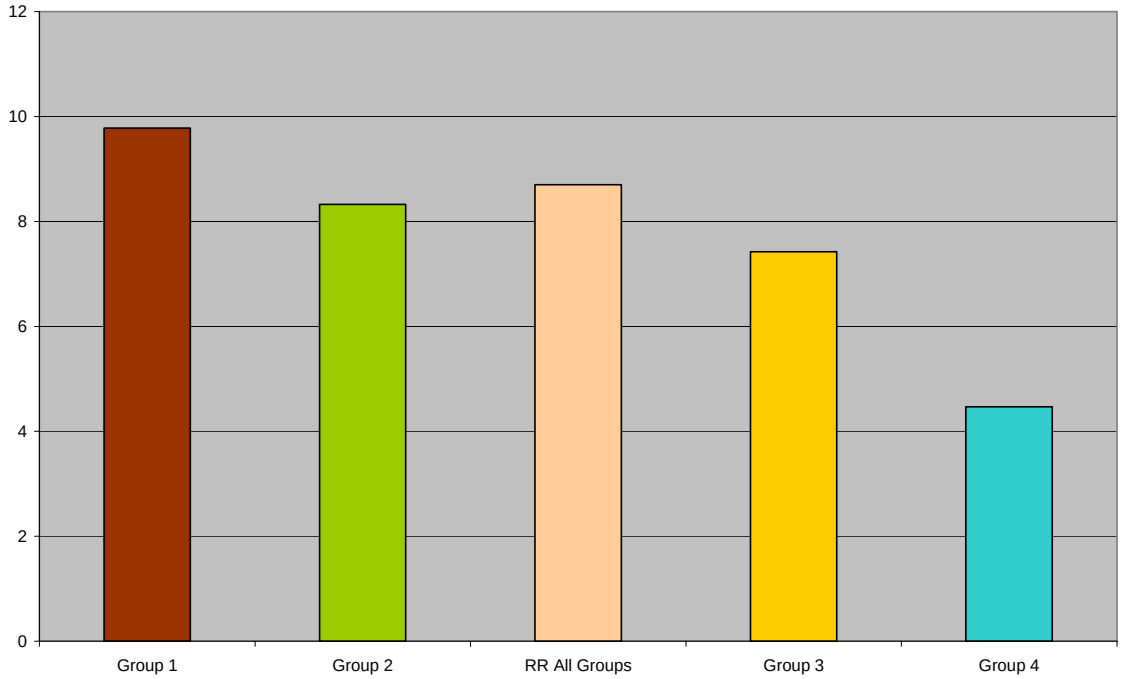

Figure 4

RH Group Contribution Averages and RR Contribution Average (Periods 16t25)

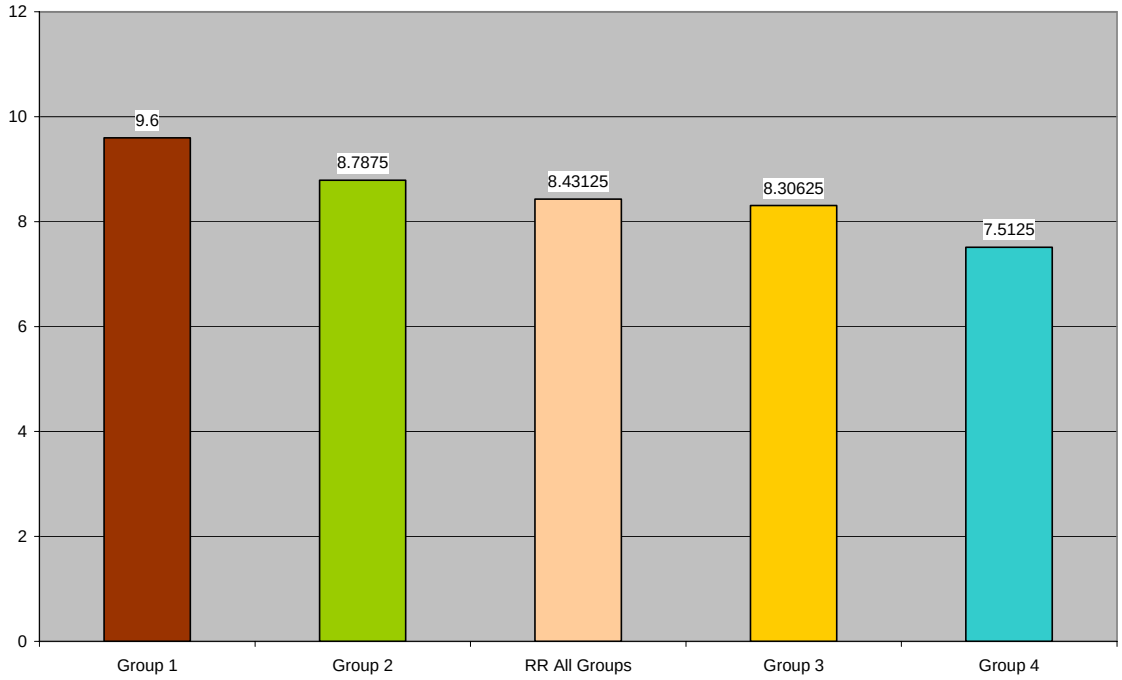

Figure 3 graphs the average levels of contributions during periods of homogeneous grouping in the HR treatment sessions. Each of the left two and right two bars represents the average contribution in a set of groups—the Group 1 bar, for example, being the average contribution in the groups of highest contributors and most vigorous non-perverse punishers in the four sessions of the treatment. As is easily seen, the ordering is exactly as would be predicted based on diagnostic period behaviors. For comparison, the middle bar represents the average contribution of treatment RR subjects during the same ten periods (6 – 15). Using Mann-Whitney tests, we find that average contributions by groups of first-ranked subjects in the HR treatment significantly exceeded those of RR treatment counterparts during the same periods (6 – 15). The HR treatment's fourth-ranked groups contributed significantly less than RR groups.[23] These results follow expectations, since first-formed groups, consisting (based on diagnostic period behaviors) of high contributors inclined to punish free riding, are expected to contribute more than randomly formed groups, which may contain not only low contributors and non-punishers but also perverse punishers. On the other hand, fourth-ranked groups, which are likely to contain some of the latter types and no high contributors or punishers of free riders, would be expected to contribute less, on average, than randomly formed groups.

Figure 4 parallels Figure 3 but graphs the average contributions by group during periods 16-25 of the RH treatment sessions and the average contributions of RR subjects during the same periods (16 – 25) of their sessions. Once again, the rank order of the contribution averages is exactly as would be predicted assuming correct type identification and persistence—this despite the fact that the inclination to contribute and to punish low contributors must in this case persist through 15 periods of contribution and punishment decisions in non-homogeneous groups. Once again, the comparison with average contributions in the RR treatment meets basic expectations, although only the dominance of Group 1 over RR contributions is significant according to a Mann-Whitney test. [24]

Consider also the lines for average contribution in periods 6-15 by the four HR groups, in Figure 2, which permits comparisons to be made period by period rather than

[23] Third-ranked groups also contributed significantly less than RR subjects, according to these tests.

[24] The Mann-Whitney tests discussed in this paragraph and the previous one take the behavior of each group of 4 averaged over ten periods as its units of observation. One test, for example, takes four observations, each being the average contribution in periods 6-15 in the group of highest-ranked subjects in one of the four HR sessions, and compares these with sixteen observerations, each being the average contribution in periods 6-15 for one of the four groups in one of the four RR sessions. This test yields a Z statistic of –1.798, which corresponds to an exact significance level of 0.08 in a 2-tailed test or 0.04 in a 1-tailed test of the hypothesis that groups of highest-ranked subjects contributed more than randomly formed groups in RR sessions. The comparison of lowest-ranked HR subjects to RR subjects gives a Z statistic of –2.979, which is significant at the .001 level in both 2- and 1-tailed tests. The *p*-values for the corresponding Mann-Whitney tests for the RH versus RR treatment groups in periods 16-25 is 0.05 in a 2-tailed (or 0.025 in a 1-tailed test) for the comparison of highest-ranked to RR subject groups, but the difference between lowest-ranked subjects and RR subjects is not statistically significant. Two other comparisons give statistically significant results broadly consistent with the ranking hypothesis, although not necessarily required by it: for the HR treatment, the groups of third-ranked subjects gave less than RR subjects, significant at the 10% level in a 1-tailed test, and for the RH treatment, the groups of second-ranked subjects gave more than RR subjects, significant at the 10% level in a 1-tailed test.

for ten period averages. When the four groups' behaviors are averaged across the four sessions, contributions by the first-ranked group substantially exceed those of the others, approaching and sometimes reaching 100% of endowment except in period 15. Correspondingly, contributions by the last-ranked group are always substantially below those of the others, averaging only around 40% of endowment in most periods. That group's sharp drop in contributions between periods 6 and 7 is interesting, as it suggests that in period 6 members expected the more heterogeneous behaviors, including punishment for free riding, that they had found in their mixed groups of periods 1 – 5, and that, upon seeing lower contributions and little punishment of free riding by new group members, they quickly shifted contributions downwards (rather than upwards, as did the heterogeneously formed groups in period 2). Contributions by the second and third-ranked groups are not as sharply differentiated from each other, but in line with what would be predicted if subjects were correctly "typed" based on their diagnostic period behaviors and if types were persistent, the average contribution in the second group exceeds that in the third in all but one period.

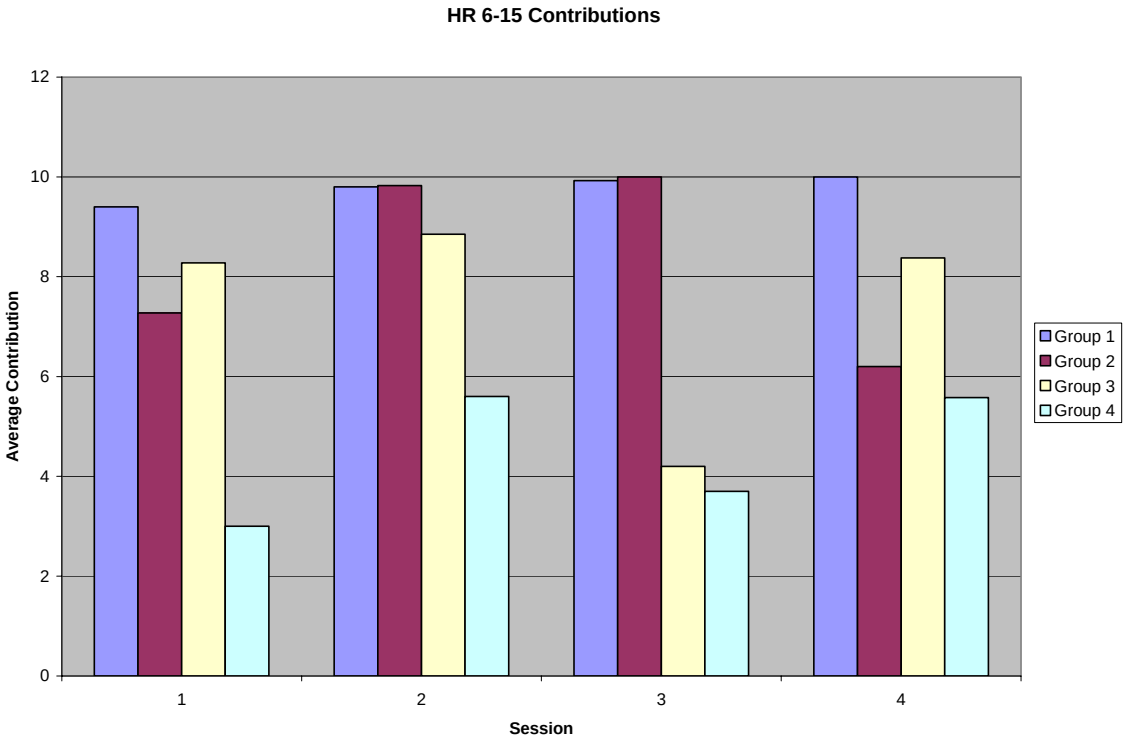

Figure 5

An especially stringent test of the ordering effect entails checking the rank order of contributions by groups 1, 2, 3 and 4 at the individual session level. In Figure 5, each cluster of four bars shows the average contributions by group in periods 6 – 15 of one of the four HR sessions. The orderings of the bars are nearly perfect in sessions 2 and 3, but show one substantial deviation from the predicted order in sessions 1 and 4. Although the ordering is not quite perfect in any individual session, it can be demonstrated that there is far too much correspondence with the predicted ordering to be reasonably attributed to chance. The probability of this close a correspondence to that predicted

occurring accidentally is about 8 in a million, on which basis we can say that the hypothesis of successful typing and persistence cannot be rejected at the 0.1% level. [25]

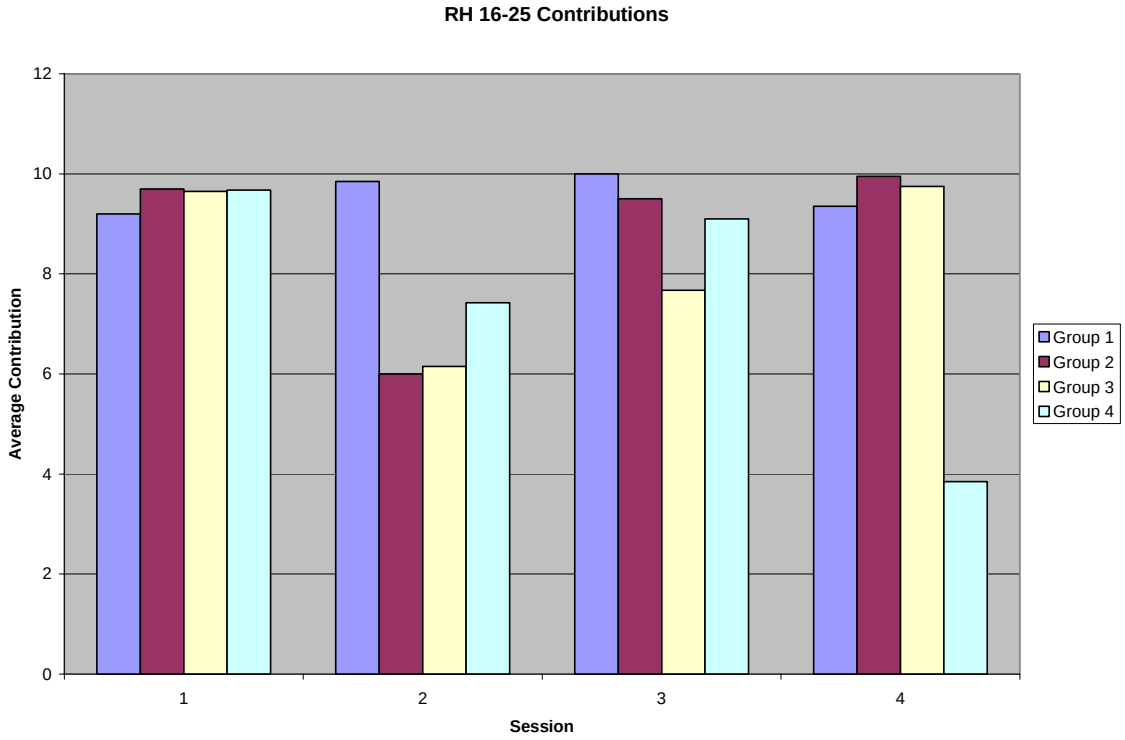

Figure 6

Figure 6, the analogue to Figure 5 for the homogenously grouped periods of the RH treatment, shows that the order of average contributions corresponds too imperfectly with the ordering by early contributions for us to reject the null hypothesis that contributions are randomly ordered within sessions.[26] It should be recalled that for types identified in periods 1 – 5 to persist into homogeneous grouping periods 16 – 25 in the RH treatment is considerably more demanding than is persistence to periods 6 – 15 in the HR treatment owing to the ten intervening periods of play in randomly formed groups. Perfect ordering is also made more challenging by the successful habit of high

[25] Formally, since there are four bars, 24 orderings of height are possible for each group. The probability that a perfect tall-to-short ordering would occur by chance on a single trial (i.e. session) is therefore 1/24. There are also three ways in which the bars could be ordered with exactly one violation of the predicted order (letting h denote height and the number in parenthesis the group formation order number, we could have h(1)>h(3)>h(2)>h(4) or h(2)>h(1)>h(3)>h(4) or h(1)>h(2)>h(4)>h(3)). Overall, then, there are four ways in 24 (p = 1/6) of having no more than one violation. In fact, Figure 4 shows no more than one violation in any of the four sessions. The probability that not more than one violation of the "perfect" ordering would occur in four of four trials if the ordering were in fact random is therefore $(1/6)^4 \cong 0.0008$.

[26] In a loose sense, there might appear to be only one exception to perfect ordering in session 1 (i.e., the first group contributed too little), only one exception in session 3 (Group 4), and only one in session 4 (Group 1). Were this correct, ordering could be called significantly different from random at conventional levels. However, the "abnormality" in session 1 must actually be counted as three exceptions, because Group 1 failed to contribute more than each of three groups it "should" have dominated. The same applies to session 4. Hence the failure to establish statistical significance.

contributions achieved in the RH sessions thanks to the presence of punishers of free riders in most diagnostic groups during periods 1 – 5 and in most random groups during periods 6 – 15.[27] The fact that at least some regularities hold—e.g., Group 1 contributions exceed those of Group 4 except in session 1, and there are "correct" orderings of three out of four groups in three of the sessions[28]—can be taken along with the aggregated results in Figures 2 and 4 as evidence that there was some persistence of type in this treatment.

5. Results, Part III: "Type" and experience at the individual level

Having shown that individuals can be identified by type and that type shows some persistence over time does not mean ruling out that different experiences will differently affect their behaviors. The roles of (initial) type and experience are now studied at the level of individual subjects.

Result 7. *Own diagnostic period contributions are a significant predictor of own contributions in later periods in the HR treatment and in periods 6-15 of the RH treatment, but they do not predict as well later period contributions in the RH treatment.*

Table 5 summarizes a set of regressions at the individual level in each of which the subject's average contribution during diagnostic periods 1-5 is the sole explanatory variable apart from a constant. The dependent varibles, from left to right, are the same subject's average contribution in periods 6-15, in period 16 alone, in periods 16-20, in periods 21-24, and in period 25. Periods 6-15 are considered together because they are played in a stable grouping just after the diagnostic periods. Period 16 is singled out as the first period under the final subject grouping, and thus as a relatively late period in which a subject's choices have not yet been influenced by his or her last set of partners.[29] Periods 16-20 are also looked at together because, while period 16's choice is less subject to the last group's influence, an average over several periods may better reflect underlying tendencies by cancelling out short-term and perhaps random fluctuations. Period 25 is of special interest as the only period in which no reputational considerations should influence subjects' choices, hence true conditional cooperators and strategy-minded mimickers of cooperation may part ways in this period.[30] Periods 21-24 are considered as late, but still potentially strategy-influenced, choices.

---

[27] To see this, compare Figures 1 and 2, then note, in Figure 6, how the ordering failures in sessions 1, 3 and 4 relate to the high contribution outcomes in almost all groups.

[28] See again footnote 26.

[29] In the HR treatment, in particular, the reinforcement of the subject's own tendencies due to playing with subjects of similar type during periods 6-15 may be influencing his or her behavior now more than it will after some periods of final-group play.

[30] The separation isn't perfect because a conditional cooperator might contribute less if he or she believes that others will not contribute in the last period, while a strategic payoff-maximizer might contribute even in the last period in our VCM-with-punishment because of the possibility of being punished by a strong negative reciprocator.

The regressions for the HR treatment subjects show that a quarter or more of the variation in own average contribution is explained by own contributions of periods 1-5 in all the periods examined except period 25, in which the effect of own initial contributions falls short of significance at the 10% level and less than 5% of the overall variance is explained, based on the R-squared statistic. Own early contribution explains more than 50% of own average contribution in periods 6-15 of this treatment, when own tendency is reinforced by play with like others. For RH treatment subjects, the explained variance falls below a quarter except for periods 6-15, and apart from those periods, the coefficient on own period 1-5 contribution is significant only once at the 10% level and only in period 16 and periods 21-24. Even these results suggest a certain persistence of type: the highly significant result for periods 6-15 occurs despite having been placed in a group of players selected for their heterogeneity during for the first five periods and then playing periods 6-15 themselves among what is probably also a heterogeneous group, randomly selected. All other coefficients are positive and of similar magnitude, except that in the period 25 regression. The marginally significant coefficient for periods 21-24 may partly reflect reinforcement by being grouped with similar types in periods 16-25.

		Ci,6t15	Ci,16	Ci,16t20	Ci,21t24	Ci,25
HR	Coeff.	0.837***	0.655 ***	0.510 ***	0.618 ***	0.375
	Std.Err	(0.11)	(0.20)	(0.16)	(0.14)	(0.23)
	R^2	0.526	0.249	0.255	0.288	0.044
RH	Coeff.	0.479 ***	0.421 *	0.308	0.396 *	-0.049
	Std.Err	(0.11)	(0.22)	(0.21)	(0.23)	(0.36)
	R^2	0.272	0.064	0.032	0.058	0.000

Reported standard errors are adjusted (Huber-White).

Table 5. Own early contributions as a predictor of own later contributions.
Note: OLS regressions with robust standard errors. Each regression included a constant, not shown here.

Result 8. *The combination of own "type" and experience with others ("environment") can account for up to three-quarters of the variation in late contribution decisions by individuals if own later contributions, which may reflect both type and environment, are included in the analysis.*

Tables 6a and 6b report series of regression equations in which a subject's own contributions in various periods are the variables to be explained, and that subject's later contributions, the contributions of other group members, and the subject's history of punishment, are included as explanatory variables. The specifications in the two tables are identical, with the first reporting results for HR subjects and the second those for RH subjects.

Each table's first column tests for whether the typing by first period decisions on which the period 2 – 5 assignment is based actually persists into those periods. In both treatments, first contribution is a significant positive predictor of period 2 – 5 average contribution at the 5% level or better despite the heterogeneous decisions of the others

with whom subjects were placed, and the average contribution of others in one's initial group shows no significant effect.

HR	Ci,2t5	Ci,6t15	Ci,16	Ci,16	Ci,16t20	Ci,16t20	Ci,21t24	Ci,21t24	Ci,25	Ci,25
Constant	7.204 ***	1.600	1.926	1.402	3.700 **	1.812	-1.607	-1.297	-1.101	-1.299
	(2.69)	(1.28)	(1.92)	(2.09)	(1.67)	(2.00)	(1.48)	(1.76)	(3.08)	(3.15)
Ci,1	0.294 **									
	(0.14)									
C-i,1	-0.286									
	(0.30)									
Ci,1t5		0.823 ***	-0.205	-0.093	-0.105	0.140	0.302 **	0.245	-0.269	-0.149
		(0.10)	(0.16)	(0.23)	(0.18)	(0.25)	(0.12)	(0.15)	(0.35)	(0.43)
C-i,1t5		0.021	-0.009	-0.052	-0.115	-0.088	-0.086	-0.084	-0.251	-0.266
		(0.17)	(0.23)	(0.21)	(0.14)	(0.14)	(0.14)	(0.14)	(0.32)	(0.33)
Ci,6t15			1.123 ***	1.196 ***	0.719 ***	0.669 ***	0.006	0.005	0.511	0.426
			(0.18)	(0.18)	(0.21)	(0.22)	(0.20)	(0.20)	(0.43)	(0.45)
C-i,6t15			-0.139	-0.218	0.048	0.013	-0.123	-0.114	-0.170	-1.440
			(0.16)	(0.15)	(0.18)	(0.18)	(0.13)	(0.12)	(0.27)	(0.27)
Ci,16t20							0.638 ***	0.616 ***		
							(0.19)	(0.21)		
C-i,16t20							0.429 ***	0.477 ***		
							(0.14)	(0.16)		
Ci,16t24									0.430	0.408
									(0.49)	(0.56)
C-i,16t24									0.657 *	0.602
									(0.39)	(0.50)
PH,1t15				-0.531		0.237				
				(0.51)		(0.29)				
PL,1t15				0.453		0.632				
				(0.45)		(0.41)				
PH,1t20								0.029		
								(0.31)		
PL,1t20								-0.232		
								(0.44)		
PH,1t24										0.865
										(0.65)
PL,1t24										0.197
										(0.96)
R^2	0.234	0.525	0.639	0.652	0.596	0.636	0.759	0.762	0.382	0.395

Table 6a. "Type" and "Environment" as Predictors of Later Contributions (HR)
Note: OLS regressions with adjusted standard errors.

Column 2's specification resembles the first column of Table 5, except that the average contributions of others' in one's period 1 and periods 2 – 5 groups are entered along with one's own average contribution in those periods to explain own average contribution during the third grouping, periods 6 – 15. For both the HR treatment, in which subjects are grouped with like others in the latter periods, and the RH treatment, in

which those periods are played in a random grouping, own 1 – 5 contribution has a highly significant positive coefficient, while the coefficient on others' contributions is not significant. The effect of environment is indirectly suggested, nevertheless, by the fact that almost twice as much of the variance in own period 6 – 15 contribution is explained in the regression for the HR treatment in which own "type" is reinforced by playing with like others as in the RH treatment, where subjects encountered random others during those periods.

RH	Ci,2t5	Ci,6t15	Ci,16	Ci,16	Ci,16t20	Ci,16t20	Ci,21t24	Ci,21t24	Ci,25	Ci,25
Constant	5.412 ***	6.471 ***	1.757	1.952	3.433	0.659	1.368	-4.418	-3.867	-16.95 *
	(1.40)	(1.65)	(2.55)	(3.80)	(4.65)	(5.69)	(2.32)	(3.06)	(7.87)	(10.09)
Ci,1	0.230 ***									
	(0.06)									
C-i,1	0.212									
	(0.13)									
Ci,1t5		0.466 ***	-0.166	-0.257	-0.193	-0.128	0.371	0.660 **	-0.420	0.239
		(0.10)	(0.20)	(0.23)	(0.18)	(0.23)	(0.23)	(0.29)	(0.42)	(0.49)
C-i,1t5		-0.169	-0.257	-0.176	-0.475	-0.373	-0.159	-0.160	0.723	0.533
		(0.21)	(0.22)	(0.24)	(0.44)	(0.45)	(0.26)	(0.25)	(0.78)	(0.77)
Ci,6t15			1.169 ***	0.959 **	0.873 ***	0.825 ***	-0.683 **	-0.340	0.380	1.131 *
			(0.36)	(0.36)	(0.28)	(0.27)	(0.27)	(0.21)	(0.65)	(0.65)
C-i,6t15			0.058	0.264	0.369	0.539	0.400 **	0.337 *	-0.246	-0.214
			(0.29)	(0.32)	(0.33)	(0.38)	(0.19)	(0.19)	(0.70)	(0.69)
Ci,16t20							0.626 ***	0.683 ***		
							(0.16)	(0.16)		
C-i,16t20							0.274 *	0.219 *		
							(0.14)	(0.13)		
Ci,16t24									0.527	0.659 *
									(0.34)	(0.37)
C-i,16t24									0.248	0.127
									(0.46)	(0.43)
PH,1t15				0.740 **		0.607 **				
				(0.31)		(0.27)				
PL,1t15				-0.599		-0.009				
				(0.59)		(0.51)				
PH,1t20								-0.470		
								(0.48)		
PL,1t20								1.426 **		
								(0.67)		
PH,1t24										-0.686
										(1.03)
PL,1t24										3.486 **
										(1.53)
R^2	0.198	0.281	0.395	0.433	0.299	0.335	0.661	0.711	0.135	0.218

Table 6b. "Type" and "Environment" as Predictors of Later Contributions (RH)
Note: OLS regressions with adjusted standard errors.

The remaining regressions are presented in pairs, the first regression in each pair being an attempt to explain own later contributions by means of own earlier contributions and earlier contributions of others in one's groups, only, while the second regression in each pair adds two variables reflecting one's experience with being a recipient of punishment. The variables labeled PH measure the average dollars of punishment received per period when contributing as much or more than the average contributed by others in one's group (PH for "punish high"), while those labeled PL measure the average dollars of punishment received per period when contributing less than the average contributed by others. These averages are calculated for periods 1 – 15, periods 1 – 20, or periods 1 – 24, always terminating before the period or periods whose contributions are being explained. Overall, at least one of the own past contribution terms is usually a significant predictor of one's later contribution(s), with more recent periods being significant more often than are earlier ones. Usually, the contributions of others in one's groups do not exert a significant effect. Only 4 of 16 coefficients on the punishment terms, all being for the RH treatment, achieve statistical significance. Two of these are of "reasonable" sign: the more the individual was punished when contributing less than his or her group-mates, the more is she contributing in the period(s) covered by the dependent variable, demonstrating the incentive effect of punishment (as shown in terms of one period changes in Table 3). But there are unexpected positive signs for being punished when a high contributor.[31]

For period 25, the overall explanatory power of the regressions are lower, as in Table 5, consistent with our remarks on end-game effects and the weakening of strategic incentives. The F-statistic for significance of the combined coefficients is in both cases significant at the 1% level, and in the regression for RH subjects two own contribution variables are significant at the 10% level. In that regression, the effect of past punishment for low contributions is also highly significant and a large positive number, indicating that although it is higher contributors who give more in the last period, *ceteris paribus*, the fear of punishment also plays a significant role. The estimate suggests that for each dollar by which punishment for free riding increased in (average) per period terms, final period contribution rose by over three dollars.

Result 9. *There was persistence of "type" with respect to punishing behavior also, in the HR treatment, but not in the RH treatment.*

To see whether subjects persisted in punishing low contributors, punishing high contributors, or not punishing, based on their displayed tendencies in early periods, we estimated OLS regressions in which the dependent variable is the average punishment index that we compute (based on the formula in Appendix A) for each of periods 16 – 25 for each individual, and the explanatory variable is the average punishment index for that same individual in periods 1 – 5. Both a regression for HR treatment subjects and one for

[31] We would have expected that having been punished when a high contributor would if anything discourage the individual from contributing more, causing these coefficients to be negative. The most likely explanation for the anomaly is that early high contributors are more likely to be high contributors later but are also more likely to have received some perverse punishment.

RH treatment subjects are shown in Table 7. For HR subjects, the average early punishment index helps to predict the average late punishment index, significant at the 5% level. This suggests that subjects who showed an early inclination to punish free riders also tended to engage in punishment in later periods, while subjects who initially refrained from punishing or who punished perversely persisted in those behaviors. For the RH subjects, there is no significant relationship between the variables. Both the relative rarity of free riding and the fact that RH subjects, but not HR subjects, were homogeneously grouped during periods 16 – 25, could help to explain this difference.[32]

Dependent Variable: Period 16 – 25 Punishment Index	**HR**	**RH**
Constant	0.195 * (0.11)	0.201 *** (0.46)
Punishment Index in Periods 1 to 5	0.307 ** (0.16)	0.009 (0.06)
Number of Observations	64	64
R^2	0.271	0.000

(*) Significant in 10 %
(**) Significant in 5 %
(***) Significant in 1 %

Table 7. Late punishment behavior as a function of early punishment behavior.
Note: OLS regressions with adjusted standard errors.

Result 10. *The two tendencies, namely to contribute or not to a public good, and to punish or not free riders or high contributors, are positively and statistically significantly related for HR subjects. There is no statistical relationship between the two tendencies in the RH treatment.*

Table 8 reports the results of regression equations in which the individual's average contribution during periods 1 – 25 is the dependent variable and the individual's average punishment index for periods 1 – 25 is an explanatory variable. The results show a significant positive relationship between the two behaviors among HR subjects and an insignificant positive relationship among RH subjects. The R^2 statistic is approximately zero in the RH regression and only .09 in the HR regression, indicating that even where the relationship is statistically significant, it explains less than 10% of the variance in contributions. Thus, our theoretical conjecture that the taste for positive reciprocity (φ)

[32] Under homogeneous grouping, first-ranked subjects who tended to be both high contributors and punishers of free riders would be grouped with other high contributors and would therefore have little free riding to punish. Last-ranked subjects who had been low contributors and non-punishers or perverse punishers would be grouped with similar subjects. Much perverse punishment seems explicable as retaliation for having been punished (see CPP), so relative lack of punishment of low contributors in last-ranked groups could lead to relative lack of perverse punishment.

and the taste for negative reciprocity (***N***) are not perfectly correlated appears to receive support from our data, suggesting the desirability of distinguishing between the two tendencies in future studies.[33]

Dependent Variable: Contribution Averages Period 1 to 25	**HR**	**RH**
Constant	7.115 *** (0.44)	8.712 *** (0.19)
Punishment Index in Periods 1 to 25	1.385 ** (0.92)	0.528 (0.46)
Number of Observations	64	64
$\mathbf{R^2}$	0.091	0.000

(*) Significant in 10 %
(**) Significant in 5 %
(***) Significant in 1 %

Table 8. Contribution behavior as a function of punishing behavior.
Note: OLS regressions with adjusted standard errors.

6. Discussion and Conclusion

This paper investigates the hypothesis that people differ in persistent ways in their inclinations to support collective action, and that the way in which a given group of people will respond to a collective action problem or social dilemma therefore depends partly on the types of people who comprise the group. Some people are more willing than others to cooperate provided they see others doing so also; some are more inclined than others to punish non-cooperation; a few are especially resistant to such punishment and actually punish cooperators. People also differ in the expectations with which they enter an interaction, and in their degree of strategic sophistication.

We tested the proposition that individuals exhibit persistent differences by seeing whether we could identify types from behaviors in similarly heterogeneous initial groups. Having made our identifications, we placed subjects in groups of seemingly like type and confirmed that tendencies to contribute more or less to a public good persisted and were reinforced by playing with like others. In particular, the rank ordering of contributions among intendedly homogeneously formed groups matched the ordering of subjects by early contributions both when subjects played homogeneously immediately after the diagnostic periods (in the HR treatment) and when they did so only considerably later in the experiment (in the RH treatment). Persistence of contribution tendency was also

[33] An important theoretical issue for biologists and social scientists is to understand the evolutionary relationship between the two tendencies. Whatever promoted an increase in the proportion of pro-social punishers in human populations would have favored propensities toward cooperation, while more cooperative populations would have reduced the relative costliness to the individual of being a pro-social punisher.

confirmed by individual-level regressions. In addition, we found that early exhibited tendencies to punish free riders, to refrain from punishment, or to (perversely) punish high contributors, were present at least to some degree after ten additional periods of play, even when grouped with others by random assignment. Punishment of free riding was no less prevalent in the final period of play, supporting the notion that it reflects a taste (negative reciprocity). The corresponding evidence from the last period regarding a taste for perverse punishment was slightly weaker, although there was strong evidence of type persistence in that all of those who punished perversely in the final period had been perverse punishers or non-punishers during the early (diagnostic) periods.

Although initial inclinations or types evidently differed among subjects in ways that persisted, experience in the series of interactions also influenced later behaviors. In individual level regression, we were able to predict three-quarters of the variation in contributions using measures of own early choices and of the actions of others in one's groups. The relative importance of the two factors differed across treatments, with others' actions being more important when groups had not been made homogeneous early on (the RH treatment). The fact that the rank ordering of average contributions across groups was significantly related to ordering by identified "type" in the HR but not in the RH treatment, when examined at the session level, is another indication that both "type" and experience affect behavior.

The "social engineering" or "organizational design" implications of our experiment should also be noticed. Although we demonstrated that one can with a reasonable degree of replicability put together groups that will significantly exceed average levels of cooperation—something one might want to do, for instance, to create a more successful business partnership or team—this came, in our homogeneous grouping periods, at the cost of also creating some extremely uncooperative groups. The significantly higher average contributions achieved in the RH and in the RR treatment[34] suggest that for better average results, low contributors and punishers of low contributors should be put together, rather than squandering the efficiency-enhancing potential of the punishers by grouping them with already cooperative types. Whether one wants to achieve pockets of excellence or as good as possible an average result depends upon the problem at hand. For some purposes, the best approach seems likely to be to constitute as many groups as possible out of a mix of strong positive reciprocators, strong negative reciprocators, and more neutral or payoff maximizing types while isolating the few strongly perverse punishers in groups that must either be treated as a "lost cause," or

[34] We carried out Mann-Whitney tests to compare contributions in the RH versus the HR treatment and in the RR versus the HR treatment. The data points are averages of contributions throughout entire sessions, with one observation per session (because subjects' behaviors within a given session are not statistically independent of one another either across periods or across groups). With four observations for each treatment, the tests showed contributions to be higher in the RH and RR treatments, significant at the 5% level in a one-tailed test of the RH versus HR treatments and at the 10% level in a one-tailed test of the RR versus the HR treatment.

policed by some external mechanism.[35] Other mechanisms for limiting the negative impact of anti-social types can also be designed.[36]

We conclude that understanding agent heterogeneity is important to understanding and improving the solution of collective action problems. Cooperation doesn't decline with time in public goods experiments because a representative agent, a payoff maximizer, learns the iterated dominant solution with experience. Instead, cooperation usually declines when there is no way to control group membership or punish free riding, because in such a situation more cooperative subjects find no other way to protect themselves from free riding. If cooperative subjects teamed up with strategic-minded payoff maximizers can exclude or punish free riders, high cooperation can be sustained at least until the final periods of interaction. In the real world, where there is rarely a known and commonly shared last period, this may be an adequate solution to many problems.

[35] In a sense, this is exactly what societies do when they consign their most anti-social individuals to prisons! For an experiment in which free-riders are ejected from the main body of subjects leading to highly cooperative performance by those who remain, see Cinyabuguma, Page and Putterman (2003).
[36] See Ertan, Page and Putterman for an experiment in which choice of rules by majority vote neutralizes the impact of the least cooperative subjects.

References

Ahn, T.K., Elinor Ostrom and James Walker, Feb. 2002, "Incorporating Motivational Heterogeneity into Game Theoretic Models of Collective Action," unpublished paper, Workshop in Political Theory and Policy Analysis, Indiana University.

Anderson, Christopher M. and Louis Putterman, 2004, "Do Non-strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism," Department of Economics Working Paper 2003-15, Brown University, revised March 2004.

Andreoni, James, 1988, "Why Free Ride? Strategies and Learning in Public Goods Experiments," *Journal of Public Economics* 37: 291-304.

Andreoni, James, "Cooperation in Public-Goods Experiments: Kindness or Confusion?" *American Economic Review* 85 (4) Sept. 1995, 891-904.

Ben-Ner, Avner and Louis Putterman, 1998, "Values and Institutions in Economic Analysis," pp. 3-72 in Ben-Ner and Putterman, eds., *Economics, Values and Organization*. New York: Cambridge University Press.

Ben-Ner, Avner and Louis Putterman, 2002, "On Some Implications of Evolutionary Psychology for the Study of Preferences and Institutions," with Avner Ben-Ner, *Journal of Economic Behavior and Organization* 43: 91-99, 2000.

Bochet, Olivier, Talbot Page and Louis Putterman, forthcoming, "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior and Organization* (in press).

Boyd, Robert and Peter Richerson, 1985, *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, Robert and Peter Richerson, 2002, "Group Beneficial Norms can Spread Rapidly in a Cultural Population," *Journal of Theoretical Biology* 215: 287-96.

Carpenter, Jeffrey, 2003, "The Demand for Punishment," unpublished paper, Department of Economics, Middlebury College, January.

Carpenter, Jeffrey and Peter Matthews, 2002, "Social reciprocity," Middlebury College Department of Economics Working Paper #29.

Charness, Gary and Matthew Rabin, 2002, "Understanding Social Preferences with Simple Tets," *Quarterly Journal of Economics* 117 (3): 817-69.

Cinyabuguma, Matthias, Talbot Page and Louis Putterman, 2003, "Cooperation Under the Threat of Expulsion in a Public Goods Experiment," Department of Economics Working Paper, Brown University.

Cinyabuguma, Matthias, Talbot Page and Louis Putterman, in process, "On Perverse and Second-Order Punishment in Public Goods Experiments with Punishment Opportunities," unpublished paper, Brown University.

Cosmides, Leda and John Tooby, 1989, "Evolutionary Psychology and the Generation of Culture, Part II, Case Study: A Computational Theory of Social Exchange," *Ethology and Sociobiology* 10: 51-97.

Cox, James and Daniel Friedman, 2002, "A Tractable Model of Reciprocity and Fairness," working paper, Department of Economics, University of California Santa Cruz.

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

Durham, William H., 1991. *Coevolution: Genes, Culture, and Human Diversity*. Stanford, CA: Stanford University Press.

Falk, Armin, Ernst Fehr, and Urs Fischbacher, 2001, "Driving Forces of Informal Sanctions," Working Paper No. 59, Institute for Empirical Research in Economics, University of Zurich, September.

Fehr, Ernst and Simon Gächter, 2000a, "Cooperation and Punishment," *American Economic Review* 90: 980-94.

Fehr, Ernst and Simon Gächter, 2000b, "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives* 14 (3): 159-81.

Fehr, Ernst and Simon Gächter, 2002, "Altruistic Punishment in Humans," *Nature* 415: 137-40.

Fischbacher, Urs, Simon Gächter and Ernst Fehr, 2001, "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment," *Economics Letters* 71: 397-404.

Gintis, Herbert, 2000, "Strong Reciprociy and Human Sociality," *Journal of Theoretical Biology* 206: 169-179.

Gintis, Herbert, Samuel Bowles, Robert Boyd and Ernst Fehr, eds., forthcoming, *The Moral Sentiments: Evidence, Models, and Policy*. New York: Cambridge University Press.

Gunnthorsdottir, Anna, Daniel Houser, Kevin McCabe, and Holly Ameden, 2002, "Disposition, History and Contributions in a Public Goods Experiment," unpublished manuscript, Department of Economics and Economic Science Laboratory, University of Arizona.

Guttman, Joel, 2000, "On the Evolutionary Stability of Preferences for Reciprocity," *European Journal of Political Economy*, 16: 31-50.

Guttman, Joel, 2003, "Repeated Interaction and the Evolution of Preferences for Reciprocity," *Economic Journal* 113, no. 489: 631-656.

Henrich, Joseph and Robert Boyd, 2001, "Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas," *Journal of Theoretical Biology* 208: 78-89.

Hoffman, Elizabeth, Kevin McCabe and Vernon Smith, 1998, "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology," *Economic Inquiry* 36: 335-52.

Kreps, David, Paul Milgrom, John Roberts and Robert Wilson, 1982, "Rational Cooperation in Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory* 27: 245-52.

Kurzban, Robert and Daniel Houser, 2001, "Individual Differences in Cooperation in a Circular Public Goods Game," *European Journal of Personality* 15 (S1): S37-S52.

Ledyard, John, 1995, "Public Goods: A Survey of Experimental Research," pp. 111-94 in John Kagel and Alvin Roth, eds., *Handbook of Experimental Economics*. Princeton: Princeton University Press.

McCabe, Kevin, Stephen Rassenti and Vernon Smith, 1996, "Game Theory and Reciprocity in Some Extensive Form Bargaining Games," *Proceedings of the National Academy of Science*, November, 13421-28.

Offerman, Theo, Joep Sonnemans and Arthur Schram, 1996, "Value Orientations, Expectations, and Voluntary Contributions in Public Goods," *Economic Journal* 106: 817-45.

Page, Talbot, Louis Putterman and Bulent Unel, 2003, "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency," revised version of Working Paper No. 2002-19, Department of Economics, Brown University.

Palfrey, Thomas and Prisbrey, Jeffrey, 1997, "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *American Economic Review*; 87(5): 829-46.

Putterman, Louis with Matthias Cinyabuguma, Ioannis Garos and Theodore Marr, in process, "On Perverse Punishment in Decentralized Sanction Regimes," unpublished paper, Department of Economics, Brown University.

Rabin, Matthew, 1993, "Incorporating Fairness into Game Theory and Economics," *American Economic Review* 83: 1281-1302.

Saijo, Tatsuyoshi and Hideki Nakamura, 1995, "The 'Spite' Dilemma in Voluntary Contribution Mechanism Experiments," *Journal of Conflict Resolution* 38 (3): 535-60 (Sept.).

Schelling, Thomas, 1971, "On The Ecology of Micromotives," *The Public Interest* 25: 61-98.

Sefton, Martin, Robert Shupp and James Walker, 2002, "The Effect of Rewards and Sanctions in Provision of Public Goods," Working Paper, University of Nottingham and Indiana University.

Sethi, Rajiv and E. Somanathan, 2003, "Understanding Reciprocity," *Journal of Economic Behavior and Organization* 50(1): 1-27.

Appendix A The Reduction Index

The reduction index measures the extent to which the subject engages in using the option of reducing others' earnings in a manner likely to spur higher contributions to the group account by other subjects, thus raising group efficiency. In a given period, the index is defined as

$$RI_i = R_{i1}*W_1 + R_{i2}*W_2 + R_{i3}*W_3 + R_{i4}*W_4 \qquad \text{(A1)}$$

where the *R*'s on the right side are the number of experimental dollars by which subject i reduced the earnings of the subject specified, 1 referencing the subject who made the highest contribution to the group account in that group and period, 2 referencing the subject who made the next highest contribution, etc., and where the *W*'s are weights capturing the effects on the recipient's next period contribution behavior of receiving one dollar of reductions. Since there are only four subjects in a group and subjects cannot reduce their own earnings, the R term corresponding to subject *i* herself will always be zero. If two or three subjects were tied for highest (lowest) contributor, punishment of any of them by *i* is included in R_{i1} (R_{i4}). For the *W*'s, we used the estimated coefficients of a regression equation having the same explanatory terms as the right hand side of (A1), with the *W*'s being coefficients to be estimated, and with the dependent variable being the change in the targeted subject's contribution from the period in which reduction dollars *R* were received to the next period. The data for which this regression was estimated were the reduction only (no communication) treatment of Bochet *et al*. (forthcoming). From these estimates, $W_1 = -0.5$ (because one dollar of reductions led to an average decline of \$0.50 in the recipient's contribution when the recipient was the highest contributor in her group in period t), $W_2 = W_3 = +0.33$, and $W_4 = +0.5$ (a dollar of reductions led to an average increase of \$0.50 in the recipient's contribution when the recipient was the lowest contributor in her group in period t).[37] Thus, *RI* is a larger positive number the more i reduced the earnings of low contributors, a negative number of larger absolute value the more i (exclusively) reduced the earnings of high contributors (i.e., engaged in "perverse punishment"), and so forth.

[37] A number of specifications were tested to check whether the effect of "punishment" on next contribution was influenced by the presence of a tie for the position of top or bottom contributor, and by other factors. The final regression OLS result on which the W values are based, which captures the basic trend in all of these regressions, is

	B	Std. Error	t statistic	p value
(Constant)	-0.4919	0.1387	-3.5455	0.0004
Top	**-0.5030**	0.1505	-3.3419	0.0009
Second	**0.3388**	0.2267	1.4946	0.1357
Third	**0.3426**	0.1428	2.3985	0.0169
Bottom	**0.5385**	0.0576	9.3430	0.0000

R-squared equals .2097, F statistic 28.32.

Appendix B Experiment Instructions

[Screen 1]
This is an experiment, funded by a research foundation, to study decision-making. You will be earning money in "experimental dollars" during the experiment. At the end of the experiment you will be paid in cash in real dollars (each experimental dollar is worth a real $0.05, or five cents). The amount you will earn will depend on your and others' decisions. The maximum possible earning is $32.50 (real dollars) and the minimum possible is $5. You are likely to earn an amount in between. Please make sure you understand the decision process.

[Screen 2]

Your First Decision: Assigning Money to Group and Private Accounts

The experiment consists of a number of distinct periods or rounds of decision-making. All of these periods have the same basic structure.

In each period, you will be interacting with three other participants in the experiment, to form a group of four. The other three people who are in your group at any given time will be identified to you as "B," "C," and "D." You will not know their actual identities either while making your decisions or after the experiment.

At the beginning of the period each person in your group will receive $10 (experimental dollars). Each of you must decide how to divide this amount between a group account and a personal account.

The money you assign to your personal account goes into your earnings.

An amount equal to 0.4 times the group's total assignment to the group account goes into your earnings.

Earnings = (amount in personal account) + (0.4)(total in group account)

[Screen 3]

The next four screens are set up to help you test your understanding of the experiment. For each of the screens that follow, there is a paper worksheet on your desk. Fill in the blanks in the worksheet first, then enter the information in the practice decision screen. The numbers you type in the practice screens are for practice only and will net affect your earnings from the experiment.

Practice Questions

Practice 1.

The four members of your group each have $10. Every member of your group has assigned $10 to the group account and $0 to their personal account. Fill in the amount to the right.

1) Amount you assigned to group account: $_____
2) Amount you assigned to your personal account: $_____
   (= $10 – group account assignment on line 1)
3) Total number of dollars assigned to your group account: $_____
4) Income from the group account for a member of you group: $_____
   (0.4 • group account total in line 3)
5) Your earnings after the assignment decisions: $_____
   (group account income in line 4 + personal account income in line 2)

Now, go back to the practice screen. Type in your assignment to the group account in the window and submit it to make sure your calculations are correct.

Practice 2.

The four members of your group each have $10. Every member of your group has assigned $0 to the group account and $10 to their personal account. Fill in the amount to the right.

1) Amount you assigned to group account: $_____
2) Amount you assigned to your personal account: $_____
   [= ($10) – (group account assignment on line 1)]
3) Total number of dollars assigned to your group account: $_____
4) Income from the group account for a member of you group: $_____
   (0.4 • group account total in line 3)
5) Your earnings after the assignment decisions: $_____
   (group account income in line 4 + personal account income in line 2)

Now, go back to the practice screen. Type in your contribution in the window and submit it to make sure your calculations are correct.

Practice 3.

Person 2 assigned $10 to the group account and $0 to his personal account, person 3 assigned $5 to the group account and $5 to his personal account and person 4 assigned $0 to the group account and $10 to his personal account.

Fill in the amounts at right for the above situation assuming that you assigned $5 to the group account.

1) Amount you assigned to group account: $_____
2) Amount you assigned to your personal account: $_____
(= $10 – group account assignment on line 1)
3) Total number of dollars assigned to your group account: $_____
4) Income from the group account for a member of you group: $_____
(0.4 • group account total in line 3)
5) Your earnings after the assignment decisions: $_____
(group account income in line 4 + personal account income in line 2)

Now, go back to the practice screen. Type in your contribution in the window and submit it to make sure your calculations are correct.

Practice 4.

Person 2 assigned $10 to the group account and $0 to his personal account, person 3 assigned $5 to the group account and $5 to his personal account and person 4 assigned $0 to the group account and $10 to his personal account.

Fill in the amounts at right for the above situation assuming that you assigned $6 to the group account.

1) Amount you assigned to group account: $_____
2) Amount you assigned to your personal account: $_____
(= $10 – group account assignment on line 1)
3) Total number of dollars assigned to your group account: $_____
4) Income from the group account for a member of you group: $_____
(0.4 • group account total in line 3)
5) Your earnings after the assignment decisions: $_____
(group account income in line 4 + personal account income in line 2)

How does this change affect the earnings of other members of your group, assuming that the switch of $1 from your individual to your group account is the only change?

Now, go back to the practice screen. Type in your contribution in the window and submit it to make sure your calculations are correct.

[Screen 4]

Your Second Decision: Reductions

There is another decision that affects your earnings. Once you learn the others' assignments to the group account, you have a chance to reduce others' earnings, and others have a chance to reduce your earnings. Suppose, in the last example, you decide to reduce B's earnings by $2, C's earnings by $3, and D's earnings by $4. The total amount of reductions you make on others' earnings is $9.

It costs you $0.25 for each $1 you reduce others' earnings. So your own earnings are reduced by (0.25)($9) = $2.25 in this example.

Just as you can reduce others' earnings, others can reduce yours. Suppose B reduces your earnings by $2, C by $1 and D by $0. The total reduction of your earnings by others is ($2 + $1 + $0) = $3. You will learn that your earnings have been reduced by a total of $3 but you will not learn who has reduced your earnings by what amount.

Similarly none of the others will learn by how much you have reduced their earnings. They will only learn their total reductions by others in the group as a whole.

Please fill in the sheet labeled practice 5 and the corresponding practice decision screen.

Practice 5.

You assigned $5 to the group account and $5 to your personal account,
person 2 assigned $10 to the group account and $0 to his personal account,
person 3 assigned $5 to the group account and $5 to his personal account, and
person 4 assigned $0 to the group account and $10 to his personal account.

You reduce person 2's earnings by $2,
person 3's by $3, and
person 4's by $4.

You receive a total of $3 in reductions from other members of your group.

1. Amount you assigned to group account: $_____
2. Amount you assigned to your personal account: $_____
   (= $10 – group account assignment on line 1)
3. Total number of dollars assigned to your group account: $_____
4. Income from the group account for a member of you group: $_____
   (0.4 • group account total in line 3)
5. Your earnings after the assignment decisions: $_____
   (group account income in line 4 + personal account income in line 2)
6. You reduced the earnings of others in your group by a total of: $_____
7. This cost you: $_____
   (0.25 * the sum of your reductions from line 6)

8. Other members of your group reduced your earnings by: $_____
9. The total change in your earnings from the reduction decisions -$_____
(line 7 + line 8)
10. Your total earnings for this period: $_____

Now, go back to the practice screen. Enter and submit your reductions to make sure your calculations are correct.

[Screen 9]

Your Net Earnings

Your net earnings for a period will be:

Amount in personal account
+ (0.4)(Total in group account)
- (0.25)(Total of your reductions of others)
- total of reductions of your earnings made by others.

If this results in a negative number in any period, your earnings for that period will be set to zero.

Each period you begin with a new $10 to allocate, and each period's earnings are independent of the others.

[Screen 10]

The Experiment as a Whole

The experiment will include a total of twenty-five periods of play. Each period consists of an assignment decision by each group member followed by a reduction decision by each group member. The nature of the decisions to be made, the amount of money at stake, and all other aspects of the structure of decision-making are as has been described in the previous instruction screens, and are the same in all twenty-five periods. There are no other decisions to make in the experiment.

The twenty-five periods are grouped into three sets of periods. Right now, we will complete the instructions for the first five periods. Additional instructions for the remaining twenty periods will follow later.

The First Five Periods

Once these instructions have been completed, you will make a series of assignment and reduction decisions for five consecutive periods, then await further instructions. In the first period, you will be assigned to a group with three other subjects, as explained earlier. Who the other three persons in your group are may change after period 1. During the rest of the first five periods (i.e. periods 2, 3, 4 and 5), the membership of your group will <u>not</u>

change again. Even when group membership is fixed, the letters identifying people (B, C and D) will change randomly. Although anyone who is in your group in periods 3, 4 and 5 is also in your group in period 2, the subject labeled "B" (for example) in one period may be labeled "C" or "D" in the next period.

Total earnings

After all twenty-five periods have been completed, your net earnings will be totaled and converted from experimental dollars to real dollars. Then $5 will be added for your participation. You will receive your earnings in cash before leaving the experiment.

Conclusion

During the experiment, there is to be no talking and no communication of any kind, except for the information that is transmitted by the computers. Since you won't have a chance to ask questions later, it is important to make sure that you fully understand these instructions now. Please raise your hand now if you have any questions. The experiment will begin as soon as all questions have been answered.

[Instructions Read Aloud by Experimenter]

You will now make ten consecutive sets of assignment and reduction decisions in a newly formed group. Throughout these ten periods, you will be interacting with the same three people, some or all of whom are likely to be different from the ones with whom you played in the first five periods. Again, you will not know the identities of the other group members, since they will be identified to you only as "B," "C," or "D." Again, the person identified to you as "D" in one period may be identified as "B" or "C" in the next.

[Instructions Read Aloud by Experimenter]

You will now make ten more consecutive sets of assignment and reduction decisions in another newly formed group. As before, the people in this group are likely to be different from the ones with whom you played in the first five periods and in the next ten periods, although some could be the same. As before, you will be interacting throughout these ten periods with the same three people. As before, the identifying letters "B," "C," and "D" will change randomly from period to period.