Cooperation, Reciprocity, and the Collective Action Heuristic

Mark Lubell
Department of Political Science
Florida State University
Bellamy 542
Tallahassee, FL 32306-2230
Email: mlubell@garnet.acns.fsu.edu

John Scholz
Department of Political Science
SUNY at Stony Brook
Stony Brook, NY 11794-4392
Email: jscholz@notes.cc.sunysb.edu

Abstract

In laboratory experiments, we manipulate the levels of niceness and reciprocity of 7 simulated players in 8-person, iterated social dilemmas. We confirm that subjects cooperate at the highest rates when the simulated players produce a nice, reciprocal strategic environment. However, subjects systematically deviate from optimal responses in intermediate environments that are either nice or reciprocal, but not both.

The collective action heuristic-- a simple model of the subject's decision process based on introspection and surprise-driven search-- explains several observed asymmetries of behavior that have important implications for the evolution of cooperation and the theory of social capital:

- On average, initial cooperators gain a cooperators' advantage over initial defectors due to defectors' inability to take advantage of reciprocal environments.
- Past experience with reciprocity reduces exploitation even when reciprocity is currently absent, while past experience with non-reciprocity does not hamper cooperation when reciprocity is currently present.
- Institutions that punish non-cooperation enhance cooperation by initial defectors, but reduce cooperation by initial cooperators.

***This research was supported by NSF Grant #SBR9515344. We would like to think Rick Wilson for his helpful suggestions on earlier drafts of this paper.

Following the research of Axelrod (1984), Putnam (1993), Taylor (1987), Trivers (1971), and other contributors to the theory of cooperation and social capital, we investigate the role of two characteristics-- reciprocity and niceness-- generally associated with the development and maintenance of cooperative solutions to social dilemmas. We respond to Ostrom's (1998) request to develop a behavioral theory of collective action by exploring the behavioral relevance of reciprocity and niceness in explaining cooperation from subjects in laboratory collective action experiments. In addition, we examine the interaction between collective action strategies, past experience, and institutions.

Our approach to analyzing behavior in collective action dilemmas reflects two general assumptions. First, the collective action strategies of citizens and of experimental subjects are best understood in terms of cognitive heuristics that generate them. In a trivial sense, heuristics can generate any strategy simply by replicating the algorithm that defines the strategy. However, heuristics available to a citizen will reflect the tradeoff between the potential benefits of a repertoire of more complex strategies and the informational and decision-making costs required to apply such a repertoire to appropriate situations (Payne, Bettman and Johnson 1993). The greater the cognitive costs involved in coping with complex strategic situations, the greater the likelihood that heuristics will produce behavior that differs from predictions based on the analysis of optimal responses. The strategic complexities of collective action dilemmas suggest that such differences will be quite significant, and therefore that optimal responses will be poor predictors of behavior. Thus, the set of heuristics used by citizens will have a profound effect on a society's ability to obtain the potential benefits of cooperation.

Second, following the literature on evolutionary psychology (Barkow, Cosmides and Tooby 1992; Caporeal et al. 1989), the set of heuristics in a given society represent specialized cognitive mechanisms for solving social dilemma problems, which are an ancient and central part of human society. To the extent that they play an adaptive role in maintaining the fitness of social groups facing recurring collective dilemmas, heuristics are likely to evolve as specialized components of cognitive architecture. These heuristics have been spawned during the genetic evolution of cognitive structures (Cosmides and Tooby 1994; Frank 1988), shaped by a culture's historical experience as transmitted through socialization

(Boyd and Richerson 1985; Fukuyama 1995; Putnam 1993), and sharpened through personal experience in the cauldron of family and small-group confrontations (Coleman 1988, 1990; Hardin 1982, 1991). Our evidence suggests the heuristics are biased in favor of cooperation: they gain some of the potential advantage of reciprocity while protecting against exploitation.

We propose a behavioral model of decision-making in collective action situations and test the model using experimental social dilemmas in the tradition of Orbell, van de Kragt, and Dawes (1988), Komorita, Hilty, and Parks (1991), Komorita, Parks and Hulbert (1992), Komorita and Parks (1994), and Ostrom, Walker, and Gardner (1994). The *collective action heuristic* combines the introspection heuristic proposed by Orbell and Dawes (1991) with the bounded rationality search process of Cyert and March (1963). The laboratory experiments use computer-simulated partners to probe how individual human subjects respond to different levels of niceness and reciprocity in the simulated strategic environment, and how these responses are influenced by past experience and by institutions designed to enhance cooperation. The three main experiments test hypotheses in a repeated 8-person public good dilemma designed to replicate low-information conditions involving little institutional and structural support for cooperation. Can the niceness and reciprocity of the seven simulated players who make up the strategic environment induce subjects (the eighth player) to cooperate in such spartan conditions? Is the optimal response model implicit in most game theoretic analyses sufficient to predict subject behavior, or does the collective action heuristic provide a better predictor and hence a better basis for analyzing collective action?

Are Niceness and Reciprocity Sufficient for Cooperation?

Axelrod's (1984, see also Axelrod and Dion 1988; Axelrod and Hamilton 1981) seminal work on 2-person prisoners' dilemma tournaments concluded that niceness and reciprocity are essential characteristics of strategies capable of sustaining social cooperation. Nice strategies are never the first to defect, so they always gain the advantages of mutual cooperation when they meet other nice strategies. Reciprocal strategies punish defection by changing the likelihood of future cooperation in response to the current defections of others, thereby reducing their vulnerability to exploitative strategies.

In iterated 2-person prisoners dilemmas, for example, tit-for-tat (TFT-- a strategy that cooperates in the first round and then reciprocates the opponent's play from the prior round) punishes defectors sufficiently to ensure that cooperation is the optimal response for opponents who are sufficiently concerned about future payoffs (Axelrod 1984). Similarly, nice strategies that retaliate in response to apparent defections can sustain cooperation in a much broader set of dilemmas (e.g., Kreps 1990). Given the evolutionary advantage of TFT, it is tempting to argue that a society dominated by nice, reciprocal citizens could evolve over time. In such a society, citizens would habitually choose nice, reciprocal strategies, which would ensure that cooperation would be the individually optimal choice.

Bendor and Swistak (1997) clarified some limitations on this evolutionary argument. They first demonstrated formally that almost-nice and almost-retaliatory strategies are the most robust of all possible strategies at gaining the benefits of cooperation in a broad class of two-person evolutionary games. Strategies with both characteristics require the minimal stabilizing frequency (50%) in the population to ensure survival against any invading set of strategies. Thus, some form of niceness and (retaliatory) reciprocity appear to be the best bets for characteristics of citizens capable of maintaining cooperation. On the other hand, cooperative strategies can only guarantee survival if they comprise a majority of the population, weakening the argument that societies might spontaneously evolve a citizenry whose dominant collective action heuristics would generate nice, reciprocating strategies. This, in turn, underscores the need for empirical evidence about the relevance of niceness and reciprocity for the actual behavior of citizens.

The large body of experimental literature on social dilemmas (see esp. the reviews by Komorita and Parks 1994; Ledyard 1995) provides some empirical validation that the theoretical advantages of reciprocating strategies influence cooperative behavior in 2-person experimental games. Komorita, Hilty and Parks (1991), Oskamp (1971), Patchen (1987), Pruitt (1968), and Wilson (1973) use experimentally "simulated" reciprocal strategies to show that subjects cooperate at higher levels when playing against reciprocal strategies than against other strategies such as all-cooperate and all-defect. Subjects cooperate more frequently with a strategy offering immediate forgiveness once a defecting subject returns to cooperation. Furthermore, chronic defection is less likely against strategies that retaliate immediately, confirming the behavioral relevance of Bendor and Swistak's (1997) emphasis on retaliation. Finally, Fehr, Gächter, and Kirchsteiger's experiments (1996, 1997; see Fehr and Gächter 1998 for a review) find indirect evidence that the possibility of reciprocity increases the gains from trade between worker and firm in experimental gift exchange games. Workers respond to firms' high wage offers with a higher effort level, even though the effort level cannot be enforced *ex ante*.

However, these theoretical and empirical findings based on 2-person dilemmas are not necessarily relevant for the larger n-person dilemmas more typical of governance problems. The transparency advantage of TFT strategies in 2-person games (Axelrod 1984) disappears as the responses of multiple players provide more ambiguous evidence about the strategies of other players. Furthermore, retaliatory defection that punishes only the prior defector in 2-person TFT strategies will inadvertently punish others who continue to cooperate in n-person games. Immediate retaliation to any single defection may therefore trigger unproductive cascades of future retaliation, particularly when played against patient TFT-like strategies whose retaliation is triggered when different numbers of others defected in the previous rounds. It is therefore not surprising that simple TFT-like strategies no longer dominate simulated tournaments even in the slightly more complex strategic environment of 3-person repeated games (Fader and Hauser 1988). Although one experiment found that the proportion of players using nice, reciprocal strategies increases cooperation even in 3- and 5-person dilemmas (Komorita, Parks, and Hulbert 1992), generalizing the 2-person findings to broader social phenomena requires more than this single test.

Heuristic Models of Collective Action Behavior

As strategic complexity increases and reciprocity becomes more difficult to detect in larger repeated games, the heuristic processes subjects use to cope with this complexity may become more important than the theoretical advantages of strategies in understanding cooperative behavior. In this section we discuss three models of decision-making in social dilemma situations. The *optimal response* model reflects the heroic assumption implicit in the full-information benchmark model of game theory that citizens can discover the optimal response to a given strategic environment during the play of a repeated game, so the discovery process itself is unnecessary to analyze. The other two models reflect assumptions about the heuristic process used to analyze strategic environments. At the opposite extreme from the optimal response model, the *myopic* model assumes that citizens respond only to the initial behavior of other players, which is easily observed, and simply ignore the bewilderingly complex patterns of play that may emerge after the first round. The *collective action heuristic* incorporates a boundedly rational learning process that emulates the optimal response model under some conditions, but emulates the myopic model under others.

As we will see, each model predicts very different responses to strategic environments that differ in the frequency of nice, reciprocal strategies. If the optimal response model predicts behavior adequately, then experiments and heuristic analyses can safely be ignored in favor of theoretic analyses of strategies and evolutionary processes. However, if the simple myopic or collective action heuristics are better predictors of cooperative behavior, then theoretical analyses of the evolution of cooperation must incorporate more realistic decision-making models to be behaviorally relevant.

The Strategic Environments

We test the ability of these models to predict subject behavior in repeated 8-person public goods games. This small group setting is sufficiently large to replicate the strategic complexity of large games, but sufficiently small to possibly observe reciprocity. It tests our previous speculation (Scholz and Lubell 1998) that intense experiences in such small group settings develop the capacity for reciprocity that we found in the collective action heuristics of typical taxpayers. Our experiments will test whether reciprocity matters for average subjects even in anonymous small group settings that lack the face-to-face interactions and communication features that are known to enhance cooperation (Ostrom 1998).

Our initial experiment manipulates two dimensions of the strategic environment by varying the niceness and reciprocity of the 7 players in the subject's strategic environment. The reciprocity dimension contrasts a reciprocal environment in which cooperation is the subject's optimal response with a non-reciprocal environment in which defection is the optimal response. Reciprocal environments add one additional cooperator (up to the maximum of 7) in the next round whenever the subject cooperates, and subtract one cooperator in the next round whenever the subject defects.[1] This could represent, for example, a strategic environment in which each player has a different threshold number of players required to cooperate in the previous round before that player cooperates in the current round. Then the subject's decision triggers the defect/cooperate decision in the "marginal" player whose threshold is affected, producing the level of responsiveness simulated in our experiment.

The niceness dimension is represented by the number of players who cooperate in the first round, ranging from two of the seven simulated players in the nasty condition to five in the nice condition. To make the repeated game more realistic in non-reciprocal environments, the initial number of cooperators is randomly increased or decreased by one cooperator (e.g., the nice, non-reciprocal environment begins with five cooperators and randomly fluctuates between four and six thereafter).

These clear, simple, subject-centered operationalizations of niceness and reciprocity are intended to represent the critical aspects of these concepts, not to "represent" the universe of nice, reciprocal strategic environments or reflect their most frequent patterns in society. We could find no theoretical, simulation, or empirical explorations of the universe of strategies used in N-person social dilemmas to guide our operationalization of different strategic environments. Hence, we pragmatically selected a simple representation of reciprocity in which the level of cooperation of the entire group responds to the single experimental subject. This environment may be more transparently responsive than those generally encountered in noisier, real-world conditions. As our results will show, however, detecting responsiveness is sufficiently challenging to provide a useful test within the constraints of laboratory experiments, where less transparency in responsiveness would increase in the number of subjects required to produce significant results.[2]

In sum, we observe the level of cooperation by subjects in four strategic environments. The *nasty, non-reciprocal* environment portrays the Hobbesian society of rational fools incapable of resolving collective action problems without the intervention of government. The *nice, reciprocal* environment portrays John Stuart Mills' self-enlightened society of citizens capable of resolving collective action problems through their own behavior-- a society rich in Putnam's (1993) social capital.

The other two environments portray intermediate levels of social capital that provide critical tests about the ability of niceness and reciprocity to induce cooperation. The *nice, non-reciprocal* environment portrays a society of altruists who could be exploited by rational players. The altruistic society tests the willingness of subjects to exploit others despite the predominance of initially nice strategies. The *nasty, reciprocal* environment portrays a society of skeptical cooperators who are initially nasty, but are also willing to reciprocate. This environment of skeptical cooperators tests the ability of subjects to explore an initially nasty environment and adopt the optimal strategy of cooperation. We will refer to these intermediate environments as *altruist* and *exploration* environments, respectively.

Predicting Behavior: Optimal Response versus the Myopic Heuristic Model

Of our three models, consider first the baseline optimal response model, which assumes that subjects will discover the presence or absence of reciprocity regardless of the initial level of cooperation. If subjects are responding optimally, they quickly ignore the irrelevant dimension of niceness, cooperate when reciprocity is detected, and defect when it is not detected. Their behavior is individually rational; they protect themselves from exploitation, but are equally willing to exploit altruists.

In contrast, consider the myopic heuristic in which the likelihood of cooperation is determined solely on the most immediately available information-- the number of other players who cooperate in the first round. The first-round choice is determined randomly, and the choice in all succeeding rounds is always to cooperate if enough other players cooperate *in the first round*, and otherwise always to defect. The myopic model predicts that subjects defect in nasty environments and cooperate in nice environments, regardless of the presence of reciprocity. Thus niceness is the only relevant dimension.

Subjects using this simple heuristic would cooperate in any strategic environment dominated by nice strategies, even altruistic ones for which defection would bring higher payoffs. Like TFT, they would avoid exploitation in nasty, non-reciprocal environments dominated by always-defect strategies, since they would respond by defecting themselves once they observed the dominance of defectors in the first round. But unlike TFT, they could not gain the advantages of cooperation in exploration environments dominated by skeptical reciprocators who would defect in the first round but cooperate in response to the subject's cooperation.

Both the optimal response and myopic models support the standard expectations that a nice, reciprocal Millsian society elicits higher levels of cooperation than a nasty, non-reciprocal Hobbesian one. However, the models predict different responses to societies with intermediate levels of social capital. The myopic model predicts higher levels of cooperation in nice environments than in nasty environments, regardless of the presence of reciprocity. In contrast, the optimal response model predicts higher levels of cooperation in reciprocal environments than in non-reciprocal environments, regardless of the level of niceness.

The experiments test these distinguishing predictions. Higher cooperation in the altruist environment but not the exploration environment supports the myopic model, suggesting that altruistic societies can support cooperation, but skeptical cooperators cannot. On the other hand, higher cooperation in the exploration but not the altruist environment supports the optimal response model, suggesting that altruism alone cannot sustain cooperation, and that reciprocity is powerful enough to elicit cooperation as long as strategies are "almost-nice" (Bendor and Swistak 1997).

Bounded Rationality and the Collective Action Heuristic

Following Caporeal et al. (1989) and Barkow, Cosmides, and Tooby (1992), we suspect that the evolutionary process in human society has shaped heuristics that are more supportive of cooperation than either of the basic heuristics. We next describe a collective action heuristic that adds a bounded rationality learning process to the introspection model developed by Orbell and Dawes (1991).

Orbell and Dawes’ (1991) model explains the levels of cooperation observed in a special case of one-shot, 2-person prisoners' dilemmas in which subjects can choose whether or not to play the game. Subjects form expectations about their opponents by projecting their own choice to cooperate or defect onto the choices of the other player. Kelley and Stahelski (1970) discovered in prior experiments that cooperators and defectors differ in their expectations, with defectors misperceiving cooperators as defectors, while cooperators believe that other players have more heterogeneous motives. In the Orbell and Dawes model, cooperators are more likely to expect the other player to cooperate, and hence are more likely to choose to play. Defectors, on the other hand, expect defection and therefore generally choose not to play, given that the defect-defect payoff is inferior to not playing. Projections of cooperation based on introspection become self-fulfilling prophecies in this voluntary one-shot game setting because cooperators rarely encounter defectors, who refuse to play the game. This cooperators' advantage, if prevalent in the ecology of games that shape collective action heuristics, would therefore favor the evolution of social cooperation as suggested in Putnam's and Axelrod's work.

To apply the Dawes and Orbell model in multi-person, repeated social dilemmas, we must slightly modify the basis of introspection and add a variant of Cyert and March’s (1963) model of adaptive search. Given the critical importance of reciprocity in obtaining the benefits of cooperation, as noted in our earlier discussion of evolutionary game theory, the model assumes that *introspection provides expectations about the reciprocity of others*. Cooperators expect reciprocity. Since the optimal response to reciprocal environments in our experiment is to cooperate, the cooperator chooses cooperation in the first round and expects others to do so as well. Defectors expect no such reciprocity. Given this belief about others, defection is the optimal response, so defectors choose defection and expect others to do so. Thus cooperators and defectors differ only in their initial expectations about the reciprocity of other players. In the following experiments, we use the subject’s choice to cooperate or defect in the first round – before they have any information about the choices of others-- as a proxy to classify cooperators and defectors.

If other players behave as expected in the first round, the model continues its initial choice. Only when cooperators encounter unexpected defection and defectors encounter unexpected cooperation do they each invoke a reciprocity test procedure. If the environment proves to be reciprocal, both will cooperate; if not, both will defect. Thus, the subject’s behavior in the first round in the game reveals the subject's prior expectation about reciprocity, and the behavior of others determines whether search procedures will be invoked.

The following behavioral rules summarize the collective action heuristic:

1. **Cooperate in the first round if reciprocity is expected; otherwise defect.**
2. **Continue first-round play as long as the level of observed cooperation matches expectations; otherwise test for reciprocity.**
3. **If reciprocity is detected, cooperate for the rest of the game; otherwise, defect for the rest of the game.**

This collective action heuristic reflects the Orbell and Dawes' (1991) cognitive miser argument, which also underlies the adaptive search model. Unless confronted by surprise, the collective action heuristic allows subjects to avoid the cognitive costs involved in detecting reciprocity in complex 8-person strategic environments. Cooperators gain full cooperation in nice, reciprocal environments without invoking costly search, but still avoid exploitation in nasty, non-reciprocal environments, where initial nastiness makes them willing to pay search costs. Similarly, defectors avoid exploitation in nasty, non-reciprocal environments without invoking costly search, and eventually gain the benefits of cooperation in nice, reciprocal environments, where initial niceness makes them willing to pay search costs. The heuristic leads cooperators and defectors alike to the optimal response in both the Hobbesian and Millesian environments.

Of course, cognitive savings come at some cost in terms of optimal performance in the more intermediate environments. Cooperators will not be surprised in nice, non-reciprocal environments, and will therefore never learn to exploit. Like the myopic heuristic, they unknowingly support cooperation in altruistic settings despite the lack of reciprocity. Defectors, on the other hand, will not be surprised in nasty, reciprocal environments, and like the myopic heuristic, therefore never learn to cooperate. Thus, the collective action heuristic predicts outcomes that differ systematically from both the optimal response and the myopic models.

We test the predictions of the collective action heuristic for cooperators and defectors. First, cooperators will defect in nasty, non-reciprocal environments, and will cooperate in the other environments. Cooperators automatically cooperate in both nice environments, and learn to cooperate because of the unexpected nastiness in the nasty reciprocal environment. Second, defectors will cooperate in nice, reciprocal environments, and will defect in the other environments. Defectors automatically continue defection in both nasty environments, and learn to continue defection in the nice, nonrecipriical environment. They learn to cooperate because of the unexpected niceness only in the nice reciprocal environment. Again the intermediate environments provide the discriminating test among all three models.

Note that the collective action heuristic might lead to self-fulfilling prophecies that limit the defector's ability to learn about reciprocity. Being pessimists, defectors expect nonreciprocity, and therefore are likely to emphasize defection during the period of testing the environment for reciprocity. Defection triggers retaliation from the reciprocating environment. Some defectors then misinterpret the resulting low level of cooperation as a confirmation of their initial beliefs that other players are nonreciprocating defectors, leading to continued defection. The self-fulfilling prophecy reflects both the biased search procedures associated with bounded rationality and the biased information processing associated with "motivated reasoning" by cooperators and defectors (Kunda 1990).

Cooperators are not affected in the same way, since they are likely to emphasize cooperation in their search procedure. This bias appropriately enhances cooperation in reciprocal environments, but can have no effect on the behavior of the non-reciprocating environment. Thus bias does not prevent cooperators from learning to defect and therefore to avoid exploitation. To explore the possibility that the self-fulfilling prophecy reduces the ability of defectors to respond to reciprocity, we test the hypothesis that cooperators will decrease cooperation in the nasty, non-reciprocal environment (compared to all nice environments) more than defectors will increase cooperation in nice, reciprocal environments (compared to all nasty environments).

Is there an evolutionary argument for the existence of the collective action heuristic? Like Orbell and Dawes (1991), we base our argument primarily on the empirical tests of whether the heuristic predicts behavior. However, we also speculate that evolutionary pressures contributed to the development of specialized cognitive mechanisms required to recognize complex patterns of reciprocal behavior in group situations. The importance of reciprocity for cooperation, combined with the bewildering complexity of strategic environments that can be encountered in groups, make reciprocity-recognition a primary candidate for specialized cognitive circuits similar to those designed to detect violations of social contracts (Barkow, Cosmides, and Tooby 1992). The introspection-based cue provided by first-round play in repeated games suggests one potential trigger to bring this special reciprocity recognition procedure into play. In combination, we speculate that these specialized mechanisms produce results that are "*better than rational*: that is, they can arrive at successful outcomes that canonical general-purpose rational methods can at best not arrive at as efficiently, and more commonly cannot arrive at all." (Cosmides and Tooby 1994: 329)

In sum, the collective action heuristic predicts that behavior changes only when surprise occurs and subsequent learning confirms the need to change expectations and behavior. Cooperators will defect only when they encounter the Hobbesian world, while defectors will eventually cooperate only in the Millsian world. Cooperators will continue to cooperate in the altruist world out of ignorance, just as defectors will fail to cooperate in the nasty but reciprocal world out of ignorance. If the effects are symmetrical, the average gains and losses will not necessarily favor cooperators or defectors. If the self-fulfilling prophecy is true, on the other hand, the resultant asymmetrical effect would stack the deck in favor of cooperation. Cooperators will gain the full benefits of cooperation in reciprocal environments while still avoiding exploitation, while defectors will gain less than full benefits of cooperation.. The experimental results should shed some light on the existence and potential magnitude of this advantage.

Social Structures, Institutions, and Bounded Rationality

Given the difficulties involved in developing and maintaining cooperative solutions solely through the strategic choices of individuals, the study of collective action has focused on the critical role of social structures and governing institutions in fostering cooperative solutions. In our experiments, we investigate two scenarios that demonstrate how such social structures and institutions interact with the strategic environment and the collective action heuristic in unexpected ways.

Social Structures Shape Past Experience

The first scenario explores the hypothesis that past experience in reciprocal environments enhances cooperation. Putnam argues that the gradual accumulation of experience with reciprocity in choral societies and other voluntary organizations developed social capital in Northern Italy, and that the demise of voluntary organizations may be undermining social capital in the United States (Putnam 1993, 1995). The argument extends Axelrod's (1984) conjectures about the importance of social structures that cluster nice, reciprocating strategies in order to accelerate the evolution of cooperation. In our experimental context, the implication is that a subject's level of cooperation will be higher after exposure to nice, reciprocal environments.

Our experiments can test only the short-term impact of experiences on subjects' willingness to cooperate, which we assume reflects the adaptation of subject heuristics across games rather than the evolution of heuristics over long periods of experience. Specifically, we analyze levels of cooperation in test games that occur after two games played in one of the four initial strategic environments. We focus on how past experience shapes current behavior in the two critical test environments-- altruist and exploration-- that represent intermediate levels of social capital in which the collective action heuristic predicts that cooperators and defectors, respectively, fail to pursue their own self interest. We will first test Putnam's implicit prediction that a subject's cooperation in any environment increases after exposure to nice, reciprocal environments. Extending Putnam's prediction for our experimental settings, Hobbesian pasts will depress cooperation, Millsian pasts will increase it, and the intermediate states will bring some intermediate level of cooperation.

The optimal response model, on the other hand, would predict that experience in previous contexts would not interfere with learning in the current environment. Subjects will cooperate in the exploration environment and defect in the altruist environment due simply to the respective presence and absence of reciprocity. The myopic model might suggest that past niceness would have a minor influence compared to current niceness, so cooperation would be slightly higher among those from nice pasts and slightly lower for those from nasty pasts. Given the importance of the initial observations of the current game in the myopic model, however, cooperation will still be much lower in the nasty exploration environment than in the nice altruist environment.

Finally, the collective action heuristic assumes that past reciprocity would only affect behavior among subjects "*pre"-surprised* by the level of past niceness. A nasty past would produce a pre-surprise for cooperators, while a nice past would produce a pre-surprise for defectors. In both cases, pre-surprise should trigger search for reciprocity in the previous environment, and initial expectations can either be *reconfirmed* or *disconfirmed*. Reconfirmation should strengthen initial expectations, while disconfirmation should moderate them. Thus nasty, reciprocal environments in the past reconfirm the initial reciprocity beliefs of cooperators, leading to higher current expectations of reciprocity. Nasty non-reciprocal environments disconfirm cooperators' beliefs and lead to lower expectations of reciprocity. Similarly, the reconfirmation from nasty, non-reciprocal environments in the past lead defectors to decrease expectations of reciprocity in current games, while the disconfirmation from nice, reciprocal environments lead defectors to increase expectations. The testable hypothesis is that past reciprocity will make a difference in test game behavior for only pre-surprised subjects. *Reconfirmed* cooperators will exhibit higher levels of cooperation than *disconfirmed* cooperators in both test games. *Reconfirmed* defectors will exhibit lower levels of cooperation than *disconfirmed* defectors in both test games. But due to limitations from the self-fulfilling prophecy, pre-surprised defectors may be less responsive to past reciprocity than would pre-surprised cooperators.

Nice pasts for cooperators and nasty pasts for defectors should lead to no search, and therefore the testable hypothesis is past reciprocity should not affect the behavior unsurprised subjects in the test games. Defectors from nasty pasts should defect, and cooperators from nice pasts should cooperate, in both the exploration and the altruist test games.

Institutions Reshape Incentives to Cooperate

The second scenario explores the common condition of partial deterrence in which institutions are created to support cooperation, but the incentives fall short of the full Hobbesian power needed to counterbalance the temptation to free ride. Thus the reciprocity of the strategic environment is still critical in determining the relative advantage of cooperation versus defection for the subject. The critical question for us is whether partial deterrence enhances or diminishes the ability of reciprocity to increase cooperation.

Two theoretical perspectives lead to very different answers to this question. Since penalties reduce the temptation payoff and hence the discount factor necessary to maintain cooperation in repeated games, Axelrod's (1984) approach would suggest that even partial deterrence should enhance the likelihood that cooperation could be sustained. The greater the reduction in the temptation payoff, the more effective that TFT-like strategies should be in sustaining cooperation. We test the Axelrod hypothesis that the level of cooperation by subjects will increase with penalty, and the increase will be greatest in reciprocal environments.

The heuristic perspective, on the other hand, suggests that the imposition of external institutions and penalties may reframe social dilemmas and therefore decrease the relevance of the collective action heuristic. For example, a penalizing institution can reframe the collective action problem in a way that causes citizens to avoid the difficulty of testing for reciprocity and behaving accordingly, since they can rely instead on the penalty institution to induce cooperation (Taylor 1987). Penalties may shift decision-making modesd to self-interested calculations rather than to collective action heuristics, particularly when an intrusive monitoring system destroys other bases of cooperation (Miller 1992). Tenbrunsel and Messick (1999) found that the introduction of weak sanctioning systems actually lowered cooperation in laboratory social dilemmas designed to test this hypothesis, although strong sanctions produced the highest level of cooperation. The authors argue that the sanctioning system elicited a “business orientation” in which the size of sanction is the only thing that matters.

By extension, penalizing institutions that have positive effects in a Hobbesian world will be less effective in a Millsian world, and may actually be counterproductive. We test the hypothesis that higher punishments increase the level of cooperation in settings of low reciprocity, but the increase is smaller and may even be negative in settings of high reciprocity.

Our experiment tests the effect of partial deterrence institutions with expected punishments ranging from zero to 1/2 of the level necessary to fully deter free riding. To focus on the interaction between penalties and reciprocity we do not vary niceness in this experiment. Instead, we use an intermediate level of niceness (3 cooperators in the first round) to ensure that both cooperators and defectors have some surprise, and hence are likely to search at some modest level for reciprocity.

The self-fulfilling prophecy hypothesis suggests that the interaction of penalties and reciprocity will affect cooperators more than defectors. If cooperators achieve relatively higher levels of cooperation than defectors in reciprocal environments when penalties are not present, cooperators should also exhibit a greater drop in cooperation if penalties minimize the influence of the collective action heuristic. Higher penalties divert attention and undercut the influence of introspection, potentially leading to significant drops in levels of cooperation for cooperators. The asymmetrical hypothesis to be tested is that: penalties will have a more positive effect on the cooperation of defectors than of cooperators, particularly in reciprocal environments.

Note that the alternative models predict no interaction between penalties and reciprocity. The optimal response model would expect no effect of penalties as long as the expected penalty is insufficient to alter the advantage of defection in non-reciprocal environments. On the other hand, the myopic approach emphasizes obvious signals, so knowledge about penalties are is likely to increase expectations about cooperation, leading to increased cooperation in all environments.

Experimental Design

The 3 main experiments reported here are based on a repeated 8-person public good dilemma designed to measure the subjects' level of cooperation under different levels of niceness and reciprocity. We conduct the experiment via computer, allowing us to present the "bare-bones" of the strategic situation without the rich institutional, informational, and social contexts that often support cooperation in the real world. Subjects were told that the computer randomly linked them to 7 other students in the large groups who participated simultaneously, but each subject was actually randomly assigned to a strategic environment composed of 7 simulated players. The instructions presented to subjects are summarized in the Appendix. Subjects were drawn from the Political Science Department subject pool, and earned credit in lower-division Political Science courses for participation. Students participated in only one experiment per student.

In the baseline social dilemma used in all experiments, the subject and 7 other (computer-simulated) players are each given a hypothetical $25.00 initial endowment.[3] In every round, the subject makes a dichotomous choice to "invest" (cooperate) or "not invest" (defect) $4.00 in the group project (public good). The total contribution is then doubled and each subject receives 1/8 of the resulting amount. The net gain if all cooperate is $8-$4 = $4, and the net gain if no one cooperates is zero. As in the N-person prisoner's dilemma, defection is the dominant strategy if the game is played only once, ensuring a payoff $3 greater than cooperation regardless of the number of others who cooperate. The game is repeated for a minimum of 16 and maximum of 25 rounds, and ends with a .25 probability per round from 16-24 or is terminated at 25. Subjects were not told in advance when the game would end to minimize the end-game effects.

The strategic environment experiment to test the basic predictions of each model and social structure experiment to test the predictions about past experience are conducted simultaneously (N=224). As discussed above, the 2 (nice/nasty) X 2 (reciprocal/non-reciprocal) manipulation of the strategic environment is the basis for all experiments. Subjects play three consecutive games. The first two games have identical strategic environments, and the last game changes to either the altruist or exploration environment. The strategic environment experiment analyzes subject behavior in the first game played. The 2 (nice/nasty) X 2 (reciprocal/non-reciprocal) X 2 (exploration/altruist) social structure experiment tests H7 and H8 by looking at analyzes the effect of the strategic environment in the first two games on subject behavior in the exploration and altruism conditions of the third game.

In the partial deterrence experiments, subjects (N=144) are randomly assigned to non-reciprocal or reciprocal strategic environments, and play three consecutive games in the assigned environment. To focus on the interaction between reciprocity and deterrence, we hold the level of niceness at a middle level of 3 cooperators in all games, as noted previously. In the first game, fixed penalties are randomly assigned to each subject in amounts of $.10, $1.00, and $2.00 with 50% detection probability. The second game is the random amount experiment in which the amount of penalty is randomly assigned from a uniform range [0, $2.00] divided up into $.10 intervals, with detection probability fixed at 50%. The third game is the random probability experiment in which the probability of receiving a $3.00 penalty is randomly selected from a [0, .5] range divided into .05 intervals. ANOVA tests do not indicate any order effects, so we focus our discussion on the more informative random amount and random probability games.

Results

We analyze the percentage of cooperative choices made by subjects in rounds 2-16 as the dependent variable.[4] The first round is omitted because that choice is made before the subject receives any information relevant to the experimental condition. We use this first choice as a measure of initial predisposition to classify the subject as a cooperator or defector.

The Strategic Environment Experiment: Surprise Alters the Effects of Niceness and Reciprocity on Cooperation

The 2(reciprocity) x 2(niceness) x 2(cooperators) ANOVA analysis of the strategic environment experiment indicates significant main effects of reciprocity (F[1,215]=17.14, p<.001), niceness (F[1,215]=10.20, p<.001), and cooperators (F [1,215]=69.87, p<.001). These results are consistent with the common prediction of all models that the mean level of cooperation (averaged across cooperators and defectors) is always higher in the nice, reciprocal environment (48%) than in the nasty, non-reciprocal environment (22%).

However, ANOVA also reveals a significant 2-way interaction between cooperators and niceness (F[1, 215]=5.81, p=.02), and a less significant 2-way interaction between cooperators and reciprocity

(F[1,215]=3.11, p=.08). In addition, the significant 3-way interaction between cooperators, niceness, and reciprocity (F[1,215]=4.02, p=.05) indicates that the effects of each dimension of the strategic environment are non-additive and contingent both on the subjects' initial predisposition and the value of the other strategic dimension. The 3-way interaction rules out both the myopic model and the optimal response model, which would predict only significant main effects for niceness and for reciprocity, respectively.

To better interpret the results, Figure 1 presents the levels of cooperation observed in each of the four experimental conditions for first-round cooperators and defectors. The reciprocal and non-reciprocal environments are indicated on the horizontal axis. Lines indicate nice and nasty conditions, with dotted lines indicating the results for surprised cooperators in nasty environments and surprised defectors in nice environments.

The evidence in Figure 1 is most consistent with the collective action heuristic. The differences in slopes indicate that the effect of reciprocity differs substantially across all environments. As hypothesized by the collective action heuristic, reciprocity has its greatest effect on surprised cooperators in nasty environments (57-26=31% difference between reciprocal and non-reciprocal environments) and surprised defectors in nice environments (29-18=11% difference). The small effect on surprised defectors in comparison to surprised cooperators also supports the self-fulfilling prophecy hypothesis. The smaller difference due to reciprocity in the nice or "no surprise" conditions for cooperators (31% difference in nasty vs. 10% difference in nice) and nasty "no surprise" conditions for defectors (11% in nice vs. 5% in nasty) respectively suggests reciprocity recognition occurs primarily in strategic environments that violate initial expectations.

Niceness produces the largest difference for cooperators in the non-reciprocal condition (56-26=30% difference between nice and nasty conditions), suggesting that cooperators fail to learn about the potential for exploitation, but still retaliate against defection. On the other hand, the null impact of niceness on defectors in non-reciprocal environments shows that defectors are willing to exploit. Because they are not searching for reciprocity, a substantial proportion of both unsurprised defectors in the reciprocal condition and unsurprised cooperators in the non-reciprocal condition adopt individually non-optimal strategies.

The Social Structure Experiment: Experience Induces Asymmetric Adaptation

We next consider the impact of past niceness and reciprocity on cooperation in altruist and exploration environments encountered by subjects in the third game. On average, cooperation was slightly higher in the third game than in the comparable first-round games, so there was no observed drop-off in cooperation from repeating separate games of the sort that has been noted in long continuing games.

The 2(past reciprocity) x 2(past niceness) x 2(cooperators) x 2(test game) ANOVA analysis confirms a significant main effect for initial (first game) cooperators[5] (F[1,207]=24.05, p<.001) and past reciprocity (F[1,207]=4.67, p=.03), but not of past niceness (F[1,207]=.62, p=.43) or test game (F[1,207]=1.09, p=.31). However, significant 2-way interactions between cooperators and the test game (F[1,207]=4.07, p=.03), and between past reciprocity and the test game (F[1, 207]=12.17, p<.001) indicate that the effects of cooperators and reciprocity are different in each test game. These interactions, combined with two marginally significant 3-way interactions-- initial cooperation, past reciprocity, and past niceness (F[1,207]=3, p=.09); and past reciprocity, past niceness, and test game (F[1, 207]=3.93, p=.05)-- are more consistent with the asymmetric predictions of the collective action heuristic than the linear predictions about niceness and reciprocity based on Putnam.[6] None of the other interaction terms were significant. Given the highly significant interaction between past reciprocity and test game, we illustrate the impact of reciprocity separately in each test game in Figures 2 and 3.

The Altruist Game. The significant main effect of past reciprocity is evident in the altruist game (Figure 2), where the upward-sloping lines for all conditions support the positive effect of past reciprocity on cooperation hypothesized by Putnam. Even in this game, however, the *pre-surprised* cooperators (those from nasty pasts) and *pre-surprised* defectors (those from nice pasts) are the most responsive to the polarizing effect of past reciprocity, consistent with the collective action heuristic hypothesis that past surprise motivates learning about past reciprocity. By far the greatest difference in cooperation between non-reciprocal and reciprocal pasts occurs for cooperators surprised by nasty pasts (71%-17%=54%), with the second greatest difference for defectors surprised by nice pasts (41%-13%=28%). The large "difference in differences" (54% vs. 28%) for cooperators provides additional support for the self-fulfilling prophecy if we assume that the greater impact for cooperators is due to the greater ability of pre-surprised cooperators to detect the reciprocity in past environments.

In contrast, the difference for unsurprised cooperators (those from nice pasts) was only 57%-40%=17%, and for unsurprised defectors (those from nasty pasts) only 45%-32%=13%. Even though this responsiveness is smaller in comparison to pre-surprised cooperators and defectors, the response indicates that the effects of past reciprocity are more robust than anticipated in our model. The collective action heuristic assumed that cooperators experiencing nice pasts and defectors experiencing nasty pasts would not have searched past environments to discover the presence or absence of reciprocity, and therefore would not have been affected by it. The first refinement in our model, then, is that surprise increases learning about reciprocity, but lack of surprise does not fully suppress learning.

The extreme responses of pre-surprised cooperators and defectors are consistent with the hypothesis that initial expectations adapt systematically when tests of reciprocity are invoked. Cooperators become more cooperative and defectors less so after games reconfirming expectations about reciprocity, while the opposite occurs when expectations are disconfirmed.

To illustrate this adaptation process, compare the behavior of pre-surprised cooperators and defectors in the altruist test game with behavior the altruist baseline game (Figure 1) that took place before subjects had any previous experience. *Reconfirmed* cooperators (nasty, reciprocal past) increase cooperation to 71% in the altruist test game compared with 57% for cooperators in the altruist baseline game. Similarly, *reconfirmed* defectors decreased cooperation to 13% from 18% for defectors in the altruist baseline game. Conversely, *disconfirmed* cooperators cooperate at only 17% in the test game compared with 57% in the baseline game, strongly suggesting that cooperation depends on expectations rather than a sense of obligation to reciprocate the niceness of altruists. *Disconfirmed* defectors cooperate at 41%, well above the 18% for defectors in the baseline game, suggesting that experience with reciprocity reduces exploitation. The exploration game also supports this hypothesis, although the support is slightly complicated by the disconfirmed cooperator's search process.[7]

The Exploration Game. As suggested by the significant interaction between past reciprocity and test game, the results for the exploration test game (Figure 3) illustrate a very different relationship between past reciprocity, niceness, and the ability to discover the presence of current reciprocity. In general, the difference between cooperators and defectors is consistent with the collective action heuristic; cooperators are more likely to take advantage of reciprocity than defectors. One of the four lines in Figure 3 is still consistent with Putnam's predictions about the effects of past experience, but even more supportive of the collective action heuristic: *pre-surprised* defectors exposed to past niceness show a 14% difference due to past reciprocity, increasing cooperation levels from 21% after nice, non-reciprocal environments to 35% after nice, reciprocal environments.

However, the other three lines in Figure 3 all contradict Putnam, and two lines contradict the collective action heuristic as well. Contrary to predictions, pre-surprised cooperators from *nasty* pasts cooperated at higher levels after *non-reciprocal* (76%) than after reciprocal pasts (57%), as indicated by the downward-sloping line for nasty pasts. The nasty pasts for defectors show a similar relationship, which is contrary to Putnam and not anticipated by the collective action heuristic. The nice past for cooperators showed no impact from reciprocity, which is consistent with the collective action heuristic but contradicts Putnam.

The two unexpectedly downward-sloping lines suggest an interesting anomaly: both defectors and cooperators apparently respond to a Hobbesian past with a greater recognition of current reciprocity, and hence achieve the highest respective levels of cooperation (37% for defectors, 76% for cooperators) after experiencing the nasty, non-reciprocal past. Both Putnam and the collective action heuristic wrongly predicted these two levels of cooperation to be as low as or lower than cooperation after the nice, non-reciprocal past. We speculate that this unanticipated phenomenon might suggest a *rebound test effect* in which the frustrating experience in the hopeless Hobbesian world makes cooperators and defectors alike more willing to accept the costs of testing for reciprocity.[8] Rebound testing would presumably depend on expectations formed by subjects prior to the experiments in relatively rich cooperative environments of American universities, and the effect would presumably decline as these prior expectations were eroded.

In summary, we find that past reciprocity increases cooperation for all conditions in altruist environments, as predicted by Putnam. Thus, past experience with reciprocity reduces the need for current reciprocity to maintain cooperation in nice environments, at least in the short run. Consistent with the collective action heuristic, this effect of past reciprocity is greatest for subjects surprised by past experience, although the effect is limited for defectors by the self-fulfilling prophecy. On the other hand, past reciprocity is less important in encouraging successful exploration in nasty, reciprocal environments. Past reciprocity does increase cooperation in exploration environments for pre-surprised defectors, consistent with the collective action heuristic. Most remarkably, however, the dreadful Hobbesian environment that lacks both niceness and reciprocity appears to be the most effective for enhancing successful exploration of reciprocal environments.

This asymmetry creates a greater advantage for the evolution of cooperation than anticipated in either the Putnam or our collective action heuristic: *past experience with reciprocity can make up for a lack of current reciprocity in supporting cooperation for both cooperators and defectors, but past experience in the worst of environments does not reduce the effectiveness of current reciprocity*.

Except for the puzzling rebound effect from past Hobbesian environments, the simple collective action heuristic has proven robust in predicting that pre-surprise enhances the impact of past reciprocity. Furthermore, reconfirmation of initial beliefs generally strengthened initial expectations, while disconfirmation moderated them. The plausible rebound test effect suggests, however, that a more complex model of expectations and adjustments will be necessary to explain behavioral adaptations to the rich possibilities of experiential learning.

Experiment 3: Reciprocity and Introspection Alter the Effect of Penalties

Finally, we consider the impact of partial-deterrence institutions and reciprocity on levels of cooperation. Table 1 reports linear regression results for the penalty amount and probability experiments, with penalty representing the increasing amount of the penalty, probability representing the increasing probability of detection, and the two dummy variables reciprocity and cooperator representing respectively the presence of reciprocity (=1) and first-round cooperation (=1) by subjects. Both models are highly significant (F=9.04, p<.001; and F=12.78, p=.001), and explain 28% and 36% of the variance in cooperation for 144 subjects. The interaction terms including dummy variables allow us to analyze the impact of penalty on cooperation separately for each combination of cooperators and reciprocity (Gujurati 1995).[9]

We begin with a discussion of the penalty amount experiment, which demonstrates the most interesting relationships between the collective action heuristic, reciprocity, and partial deterrence. Figure 4 illustrates the estimated impact of penalty amounts in a style similar to the previous figures. The line representing each condition combines the effect of all relevant coefficients, and hence calculates the total estimated impact of penalties on levels of cooperation.

[Table 1 and Figure 4 about here]

The results provide clear evidence supporting the heuristic perspective over the proposition that penalties always increase cooperation. First, penalties have their greatest impact on cooperation in non-reciprocal rather than reciprocal environments: the significant *penalty*reciprocity* coefficient estimates that the presence of reciprocity reduces the overall effect of penalties by an estimated 15% for each dollar of expected penalty when compared with non-reciprocal environments. Figure 4 illustrates that punishments increase cooperation in both non-reciprocal environments, since both lines slope upward. In reciprocal environments, however, punishments increase cooperation at a reduced rate for defectors and *punishments actually decrease cooperation for cooperators.*

The decline in cooperation for cooperators in reciprocal environments supports the asymmetric impact of penalties predicted by the collective action heuristic. The significant *cooperator*penalty* coefficient corroborates this effect: each dollar of penalty increases cooperation by defectors 15% more than it increases the cooperation of cooperators. The combined effect of reciprocity (-16%) and cooperators (-15%) account for the negative impact (23.4 - 16 - 15%= -7.6% per dollar) of penalties on cooperation for cooperators in reciprocal environments.

In the ideal Millsian environments dominated by cooperators in reciprocal settings, penalty institutions apparently do more harm than good by reducing the effectiveness of reciprocity. As penalties increase for cooperators in Figure 4, the difference in cooperation between reciprocal and non-reciprocal environments drops from 40% at low penalties to less than 10% at the highest level of penalty. At that level, the decay of social capital has reduced the level of cooperation by initial cooperators to the same level achieved by initial defectors responding solely to the presence of penalties in non-reciprocal environments!

On the other hand, even weak penalties are extremely effective in the non-reciprocal Hobbesian environments, particularly for defectors. Defectors cooperate only 18% of the time as penalties approach zero, but increase cooperation to 62% at the highest penalty levels in the experiment. Penalties are also effective among cooperators in non-reciprocal environments and defectors in reciprocal environments. In these cases of intermediate levels of social capital, then, penalties provide effective support for cooperation.

The random probability experiment suggests that the asymmetrical effects of penalty amount do not fully generalize to *expected* penalty, since increasing probability of detection leads to higher levels of cooperation in all conditions.[10] Although the negative directions of the detection probability*cooperator and detection probability*reciprocity interactions are consistent with the collective action heuristic, the effects are not significant. The $3 penalty in the random probability experiment exceeds the highest penalty in the random amount experiment, and apparently is high enough to trigger the full penalty effect on cooperators regardless of the probability of detection. Since the subjects already resent the penalty, the only way to get them to cooperate is to make it obvious that free riders will be caught.

Our results do not rule out the possibility that probabilities also produce asymmetric effects at lower levels of penalty. However, other deterrence studies confirm that experimental subjects (Casey and Scholz 1991) and even business firms (Scholz and Gray 1990) respond quite differently to changes in penalty amounts than they do to probability changes producing the equivalent change in expected penalty. We suspect that the amount of penalty provides a more salient signal to the collective action heuristic than does the cognitively more challenging probability of detection (Einhorn and Hogarth 1985), and therefore has the greatest asymmetric effect on cooperators.

Conclusion

Not surprisingly, the results we report reconfirm the suggestion from evolutionary game theory that Hobbesian societies dominated by nasty, nonreciprocating citizens elicit lower levels of cooperation from experimental subjects than Millsian societies dominated by nice, reciprocating citizens. Thus, the proportion of citizens with heuristic behavior that replicate nice, reciprocal strategies seem to provide one rough indicator of a society's heuristic social capital, with the Millsian society enjoying the highest level.

However, the results also indicate that the collective action heuristic is a better predictor of behavior than the optimal response model in other strategic environments representing different levels of social capital, even in the relatively transparent reciprocal environments used in our experiments. The heuristic reduces the cognitive costs of analyzing complex strategic environments by testing for reciprocity only when the observed number of initial cooperators contradicts reciprocity-based expectations. If the innate capacity to test for reciprocity is as important as other cognitive skills required for exchange relationships that have been sharpened in the evolutionary process (Cosmides and Tooby 1994), these results should hold for the more subtle and less transparent patterns of reciprocity encountered outside our laboratory setting.

If so, several asymmetries in our experimental results that are explained by the collective action heuristic must be accounted for in a complete behavioral theory of collective action:

a) **Asymmetry in search**: Introspection and surprise-based search prevent cooperators from exploiting nice, non-reciprocal environments, but do not prevent them from cooperating in nasty, reciprocal environments. Conversely, nasty subjects learn to exploit, but fail to learn to cooperate in nasty, reciprocal environments.

b) **Asymmetry in experience**: Past experience with reciprocity enhances cooperation even when reciprocity is currently absent, while past experience with nonreciprocity does not diminish the ability of current reciprocity to enhance cooperation. This asymmetry provides an additional reason for emphasizing reciprocity as a key factor supporting cooperation.

c) **Asymmetry in institutions**: Increasing the penalties of enforcement institutions enhances cooperation in non-reciprocal environments, but actually diminishes cooperation among cooperators in reciprocal environments.

These asymmetries suggest some interesting speculations about the evolution of social capital. The first asymmetry, combined with the self-fulfilling curse afflicting defectors, can produce a "cooperators' advantage" because cooperators outperform defectors in reciprocal environments by more than enough to compensate for the advantage defectors gain from exploiting altruistic environments. If we assume all conditions in Figure 1 to be equally likely, for example, cooperators would on average earn $.75 per round more than defectors.[11] Thus, for cooperators, the collective action heuristic produces a strategy with some of the resilience associated with Tit-for-Tat in 2-person games.

Furthermore, the asymmetry in search in the collective action heuristic stabilizes both Hobbesian and Millsian societies. When new, invading strategies are encountered, cooperators who dominate Millsian societies will detect reciprocity as long as it is still present while defectors who dominate Hobbesian societies will not detect new reciprocity. These behavioral tendencies may dampen the oscillation of the vicious and virtuous cycle described by Putnam (1993). Thus, cooperation appears to be path dependent in the presence of the collective action heuristic-- difficult to create among defectors, but robust among cooperators once it is achieved.

The asymmetrical effect of experience also may enhance the stability of cooperation. Past reciprocity decreases the incidence of exploitation among both surprised defectors and cooperators in the altruist test game, as if past experience increases expectations of reciprocity in the current environment. However, past absence of reciprocity does not depress subjects' ability to learn to cooperate in the exploration test game. In fact, past Hobbesian environments seem to create a rebound effect in which subjects appear even more sensitized to the presence of reciprocity in a currently nasty environment. This suggests that at least among American subjects, the Hobbesian experimental condition violates generalized norms of reciprocity so severely that subjects begin to search for the possibility of cooperation despite the bleakness of the environment. This asymmetrical effect of past reciprocity favors cooperation by stabilizing cooperation in nice environments where exploitation is possible, while not retarding the evolution of cooperation in nasty environments where cooperation is possible.

The asymmetry in institutions suggests that penalties and reciprocity become substitutes rather than complements in the maintenance of cooperation at some level in the development of social capital. Penalties may enhance cooperation in a Hobbesian society and among defectors in an otherwise Millsian society, but they appear to reduce cooperation among the cooperators who dominate a Millsian society. From a policy perspective, these results underscore the importance of avoiding the high costs of coercive enforcement when contingent compliance already sustains law-abiding activity (Levi 1988; Scholz and Lubell 1998a, 1998b). Coercive institutions need to tailor enforcement to match the social capital in different subpopulations and jurisdictions.[12] Given the difficulties in developing such flexible enforcement institutions, we suspect that reciprocity remains the critical ingredient to preserve social capital in the heterogeneous settings of modern societies.

Finally, our findings underscore the importance of analyzing both individual characteristics and social/institutional factors simultaneously in the study of collective action. Consider, for example, the common experimental design that randomly assigns n subjects to each n-person dilemma (e.g., Ostrom, Walker and Gardner 1992). Since high penalties increase cooperation in Hobbesian environments but decrease cooperation in Millsian environments, subject pools that differ in social capital will produce different strategic environments and divergent results. By controlling for the strategic environment, institutional factors like communication that consistently enhance cooperation in experiments may be found to enhance cooperation even more in particular strategic environments, while some factors, like the penalties we investigated, may be effective in some environments but not in others.

Whether we study citizens as voters controlling the state or as subjects complying with laws of the state, our findings support integrating the established behavioral focus on individual characteristics with the new institutionalism's focus on characteristics of the system (Levi 1988). Citizen heuristics, social structures, and institutional design all provide behaviorally-relevant foundations for the theory of collective action.

Appendix: Brief Description of Instructions

We programmed the entire experiment using Visual Basic 5.0, and conducted the experiment in an on-campus computer lab where groups of up to 24 subjects participated simultaneously on individual computers. The program is available from the authors. Although the complete instructions provided by computer[13] to each subject are too long to present here, they include the following features:

1. Brief description of a group investment scenario, using housecleaning as a substantive example of a public goods problem and moving to a group investment situation. No subject discussion is allowed.
2. Mathematical details of experimental game, including presentation of payoff matrix that subjects can use for reference during game play. To minimize the potential normative bias in favor of cooperation, cooperate and defect options were described as “Invest” and “Not Invest”. Subjects are not told when the game will end, but they are told it will last a minimum of six rounds.
3. Explanatory tables describe payoffs for entire group for different numbers of investors, the returns to the individual for investing or not investing, and the effect of random market conditions on individual returns.
4. Practice exercises ask subjects to calculate their returns for investing and not investing with various numbers of other investors, and calculate the number of other investors there must be for a given return. Subjects must answer these questions correctly before moving on.
5. In the partial deterrence experiment, subjects are told that in each game there is a chance that the computer will be monitoring their group. If the computer monitors the group, the subjects are told both the probability of being detected and the amount they will be penalized if caught. The explanatory screen shows an example of how penalties affect payoffs.
6. Subjects are told that they will play three consecutive games, and will be assigned to a new group after each game.
7. The program then proceeds to the main interface, where the subjects make their investment decisions. In all experiments, subjects can see their decision history, the payoff they received in each round, and their cumulative payoff. In the penalty condition, the computer notifies subjects if they receive a penalty and also records penalty information on the screen.
8. The final screen debriefs subjects in compliance with the protocol approved by the campus Human Subjects Review Board.

Table 1. Regression Results Predicting Level of Cooperation for Penalty and Probability Experiments

Variables	*Penalty Amount Experiment*	*Penalty Probability Experiment*
Constant	16.29 (8.05)*	21.13 (6.69)**
Reciprocity	16.46 (10.06)	-2.98 (9.72)
Subject Niceness	22.59 (9.89)*	14.71 (9.30)
Niceness*Reciprocity	23.33 (9.47)*	27.74 (8.98)*
PenaltyAmount	23.37 (7.05)**	
Penalty Probability		55.63 (23.67)*
Niceness*Amount	- 15.00 (7.70)*	
Niceness*Probability		-.321(28.31)
Reciprocity*Amount	- 16.13 (7.29)*	
Reciprocity*Probability		-15.47 (27.68)
Model Fit	F=9.0**, R^2=.28; S.E.=26.38	F=12.78**, R^2=.36, S.E=25.52

Note: *p≤.05; **p≤.01. N=144. Estimates are unstandardized regression coefficients, standard errors in parentheses. Coefficients for penalty and probability experiments are reported separately in each column.

Figure 1: Effect of Niceness and Reciprocity on Cooperation

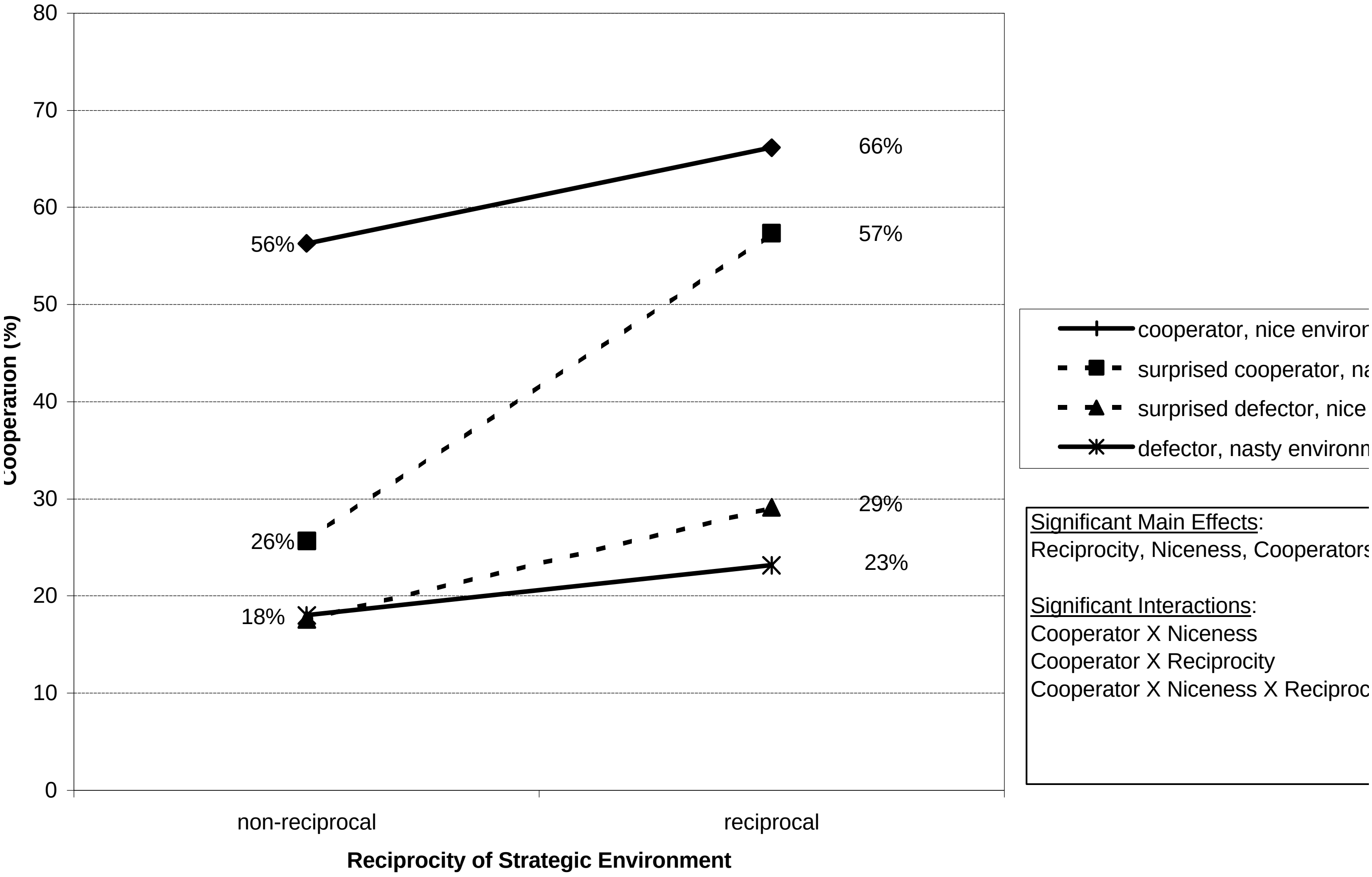


Figure 2: Effect of Past Conditions on Cooperation in Altruist Game

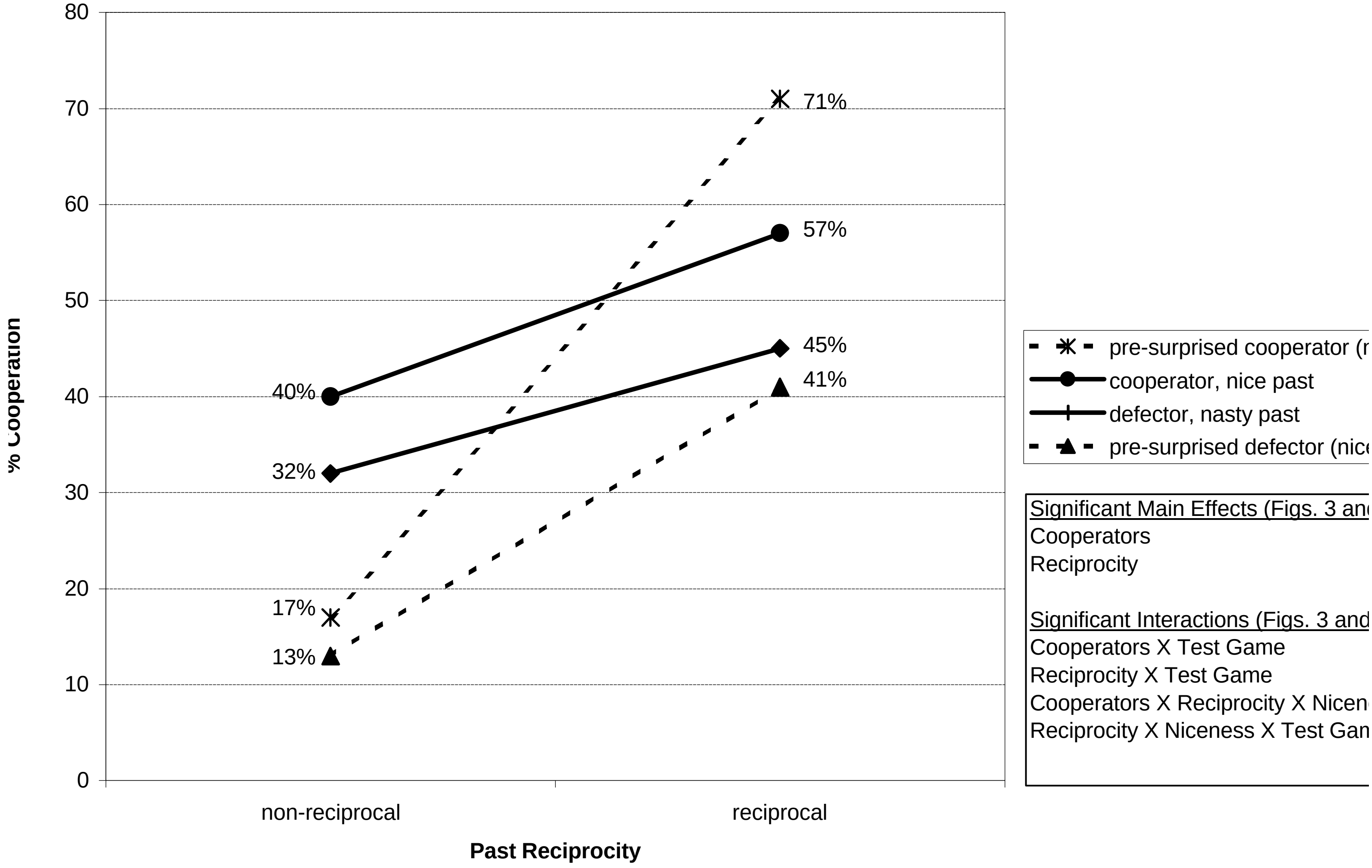
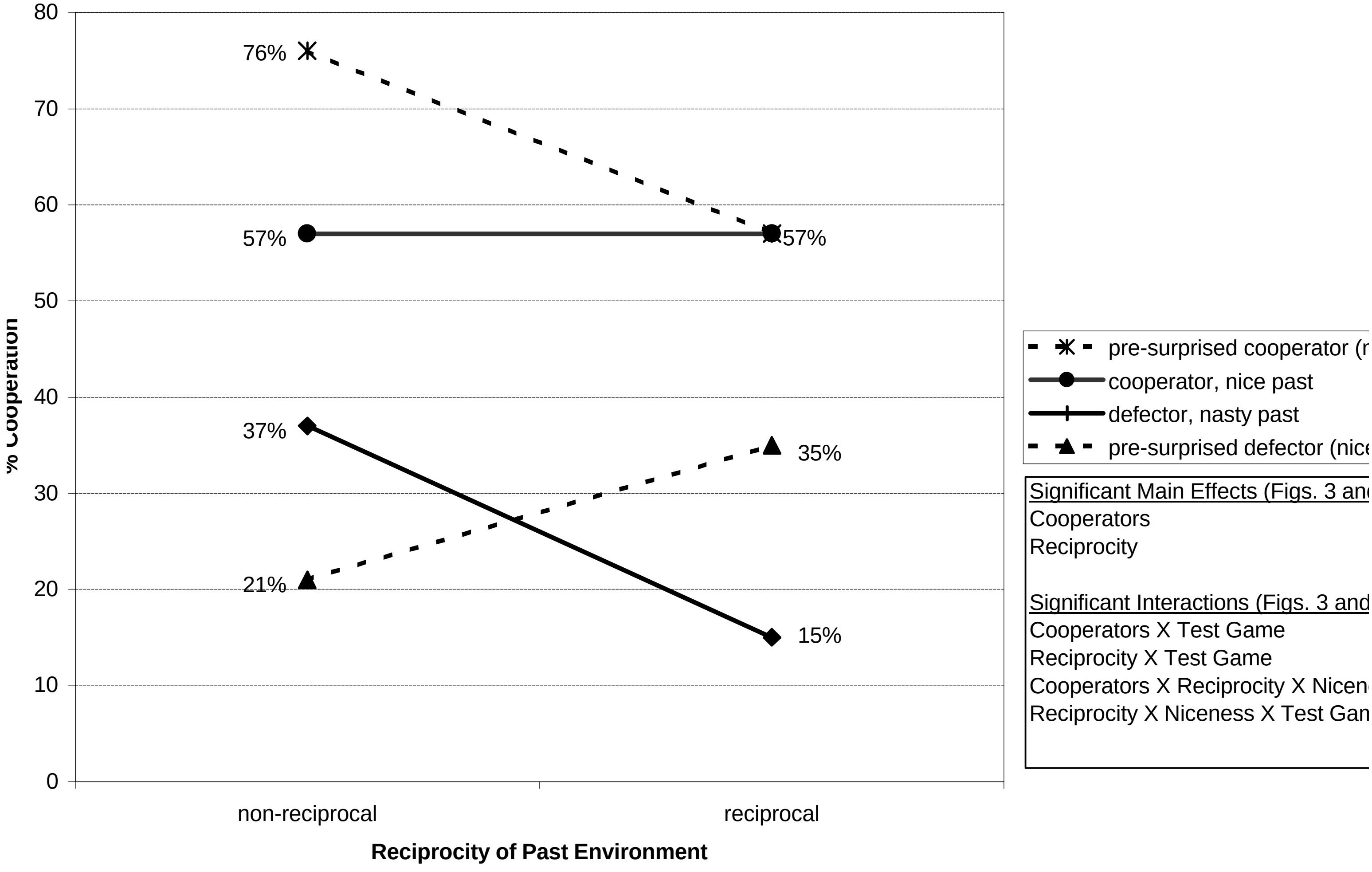
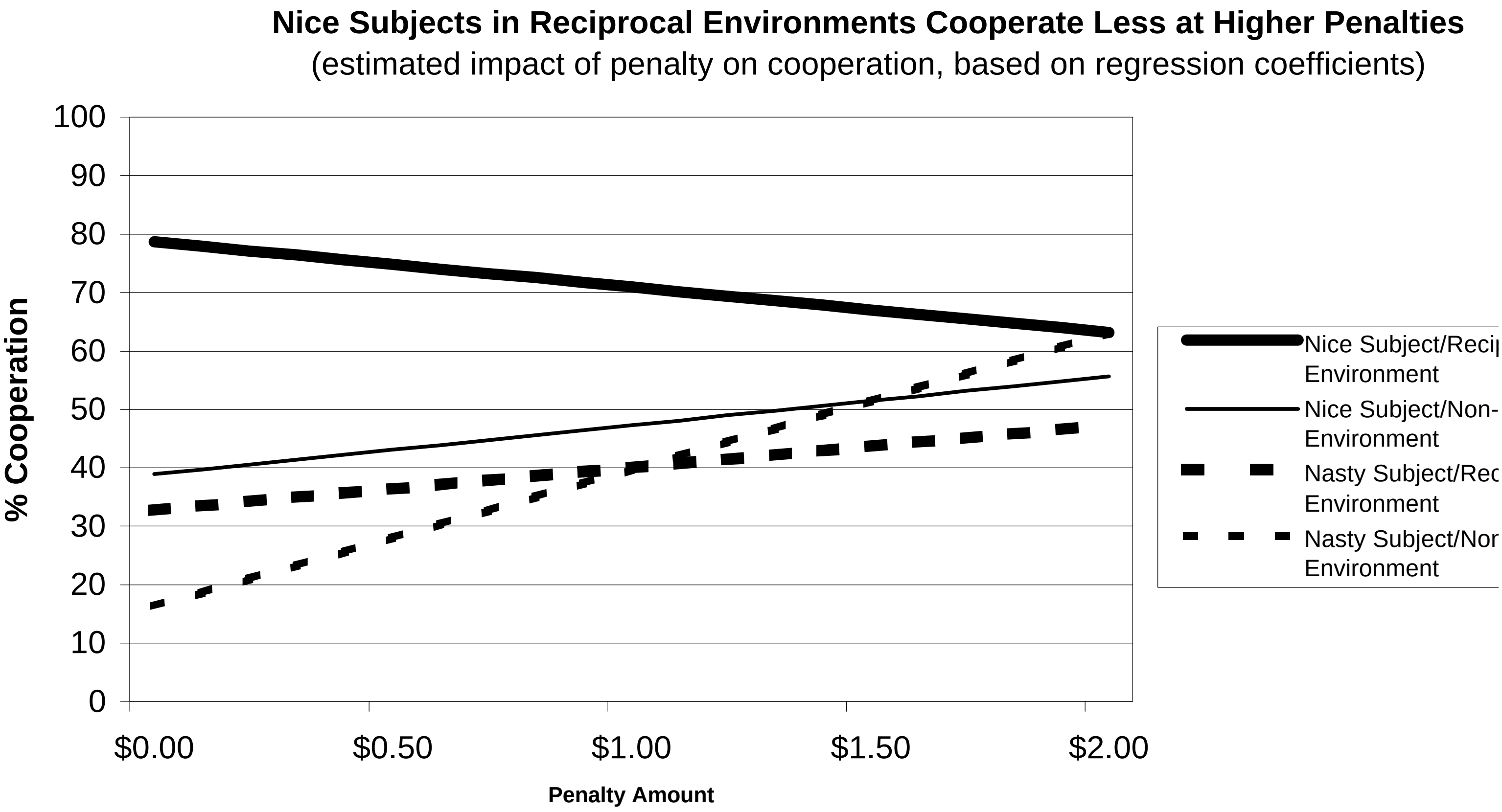


Figure 3: Effect of Past Conditions on Cooperation in Exploration Game




Nice Subjects in Reciprocal Environments Cooperate Less at Higher Penalties
(estimated impact of penalty on cooperation, based on regression coefficients)
% Cooperation
100
90
80
70
60
50
40
30
20
10
0
$0.00
$0.50
$1.00
$1.50
$2.00
Penalty Amount
Nice Subject/Reci
Environment
Nice Subject/Non-
Environment
Nasty Subject/Rec
Environment
Nasty Subject/Nor
Environment

Endnotes

[1] To ensure that full cooperation is the optimal strategy, the reciprocal environment subtracts two cooperators as opposed to the usual one cooperator whenever seven other players are cooperating and the subject defects. This prevents cycling between cooperate and defection once the ceiling of 7 cooperators has been reached from being an alternative optimal strategy.

[2] A series of experiments not reported here found, for example, that information about the individual choices of simulated players (rather than the aggregate information about payoffs) increased the impact of reciprocity on cooperation. The results were marginally significant and only tangential to the main experiments reported here, but they suggest that our relatively transparent definition of reciprocity can be made even more transparent with detailed information about individual-level behavior. The ability of subjects to detect different forms and levels of reciprocity provides an important research agenda for further explorations of the collective action heuristic.

[3] We follow the tradition of psychological rather than economic experiments in relying on hypothetical payoffs rather than real ones. Several studies demonstrate that real and large payoffs and material incentives have no effect on mean levels of behavior in experimental games (Cameron 1995; Fehr and Tougareva 1995; Knox and Douglas 1971; Komorita and Parks 1992). However, Knox and Douglas (1971) find that high stakes increase the variance in behavior across different types of subjects; defectors are more likely to defect, while cooperators are more likely to cooperate (cf. Komorita and Parks 1994, p.31). We believe that the greater behavioral differences due to high stakes increase the power of experiments to detect effects, but with enough subjects, low stakes experiments will find the same patterns.

[4] We stop at round 16 since that is the last round collected for all subjects. We could eliminate the first 5 or 10 rounds as 'learning experiences', but the number of rounds to eliminate turns out to be unimportant because they do not change the results. To avoid arbitrary cutoffs, we report the full 15 round summary measure.

[5] We use the first-game initial cooperation rather than the initial cooperation at the beginning of the test game, since the latter measure would confound the subject's expectations and behavior prior to the experiment with the impacts of the first two games on expectations and behavior.

[6] The optimal model predicts only a significant main effect for test game (reflecting the difference in current reciprocity), which was not significant. The myopic model predicts only a significant main effect of test game (reflecting the difference in current niceness), past niceness, and a possible interaction between past niceness and test game, none of which are significant. The collective action heuristic predicts a significant main effect of cooperator, an interaction between cooperator and test game, a three-way interaction between cooperator, past niceness, and past reciprocity and a four-way interaction. All but the 4-way effect are significant, but the significance of the past reciprocity, past niceness, and test game (P=.05) are not predicted. We suspect that the rebound effect on cooperators and defectors alike discussed later in the text is responsible for the unanticipated 3-way interaction and the lack of significance of the 4-way interaction.

[7] The difference for reconfirmed defectors is 21% vs 29% and for disconfirmed defectors is 35% vs 29%, although there is no difference for reconfirmed cooperators (57% vs 57%), and disconfirmed cooperators actually cooperate at greater levels (76% vs 57%). One plausible explanation is that the adaptation of expectations for cooperators is overshadowed in this condition by the reciprocity test triggered by the surprising nastiness of the exploration test game, with disconfirmed cooperators testing more energetically and successfully than others. Adaptation for defectors, on the other hand, apparently is not overshadowed by the reciprocity test in the altruist game, or the disconfirmed defectors would cooperate at the lowest level. This asymmetry between cooperators and defectors adds additional support for the self-fulfilling prophecy hypothesis, although a "rebound effect" for cooperators after Hobbesian condition that is discussed later provides an alternative possible explanation.

[8] The rebound test effect would predict that both cooperators and defectors would reach the lowest levels of cooperation in the non-reciprocal altruist game following a Hobbesian past. The prediction is confirmed for cooperators at 17%, but the 35% cooperation level of defectors is only the second lowest, not nearly as low as the reconfirmed defectors (13%) from nice, non-reciprocal pasts or even the defectors (18%) in the altruist baseline game. This increased cooperation in altruistic environments by defectors after a Hobbesian past is particularly puzzling.

[9] The 3-way *reciprocity*penalty* cooperator* interaction was not significant in an unrestricted model and introduced severe multicollinearity due to the dichotomous independent variables. Thus the interaction was dropped to focus on the remaining 2-way interactions.

[10] Remember that the maximum detection probability in the random probability experiment is .5, so the regression coefficient estimates a marginal 27% increase in cooperation for a .5 increase detection probability. This is close to the marginal effect of a $1.00 increase in the random amount experiment, where detection probability is held constant at .5.

[11] As indicated in Figure 1, cooperators make optimal choices more frequently than defectors in the reciprocal Millsian (66% vs. 29%) and exploration (57% vs. 23%) environments where cooperation is the optimal choice, for an average difference of 35.5% if both environments are equally likely. Since cooperators gain a $4 per round advantage for making the optimal decision 35.5% more often than defectors, their expected advantage is $1.44 per round. Of course, defectors make the optimal choice more frequently than cooperators in the non-reciprocal Hobbesian (82% vs. 74%) and altruist (82% vs. 44%) environments where defection is the optimal choice, for an average difference of 23% if both environments are equally likely. Since defectors gain $3 per round advantage of making the optimal defect decision, their expected advantage is $.69 per round in these two environments. Assuming that all environments are equally likely, the cooperator's advantage under the experimental conditions would amount to ($1.44-$.69)= $.75 per round. The cooperator's advantage depends on the relative advantages of cooperation over defection, of course, but would develop in most payoff schemes used in experimental research.

[12] Empirical studies find that enforcement agencies do increase the intensity of enforcement, and hence the expected penalty, in jurisdictions that have the lowest level of compliance (Mete 1999; Scholz and Wei 1986). On the other

hand, the highly valued ideal of equal treatment constrains the ability of agencies to protect social capital through such discretionary enforcement (Scholz and Wood 1998).

[13] In some of the earlier experiments, the instructions were given via paper. Later experiments included on-screen instructions.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books, Inc.

Axelrod, Robert, and William Hamilton. 1981. "The Evolution of Cooperation." *Science* 211:1390-96

Axelrod, Robert and Douglas Dion. 1988. "The Further Evolution of Cooperation." *Science* 242:1385-1390.

Barkow, Jerome H., Leda Cosmides, and John Tooby. 1992. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York, NY: Oxford University Press.

Bendor, Jonathon, and Piotr Swistak. 1997. "The Evolutionary Stability of Cooperation." *American Political Science Review* 91:290-307.

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

Cameron, L. 1995. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." Discussion paper, Princeton University.

Caporeal, Linda R., Robyn M. Dawes, John M. Orbell , and Alphons J.C. van de Kragt. 1989. "Selfishness Examined: Cooperation in the Absence of Egoistic Incentives." *Behavioral and Brain Sciences* 12:683-39.

Casey, Jeff T. and John T. Scholz. 1991. "Beyond Deterrence: Behavioral Decision Theory and Tax Compliance." *Law and Society Review* 25:821-843.

Coleman, James. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

--1988. "Social Capital in the Creation of Human Capital." *American Journal of Sociology* 94 Supplement:S95-S120.

Cosmides, Leda and John Tooby. 1994. "Better than Rational: Evolutionary Psychology and the Invisible Hand." *American Economic Association: Papers and Proceedings* 84:327-32.

Cyert, Richard M. and James G. March. 1963. *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Fader, Peter S., and John R. Hauser. 1988. "Implicit Coalitions in a Generalized Prisoner's Dilemma." *Journal of Conflict Resolution* 32:553-82.

Fehr, Ernst and Simon Gächter. 1998. "Reciprocity and Economics: The Economic Implications of *Homo Reciprocans*." *European Economic Review* 42:845-859.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econmetrica* 65:833-860.

--1996. "Reciprocal Fairness and Noncompensating Wage Differentials." *Journal of Institutional and Theoretical Economics* 152:608-640.

Fehr, Ernst and E. Togareva. 1995. "Do High Stakes Remove Reciprocal Fairness--Evidence from Russia." Discussion paper, University of Zurich.

Frank, Robert. 1988. *Passions within Reason: The Strategic Role of the Emotions*. New York: W.W.Norton and Company.

Fukuyama, Francis. 1995. *Trust: The Social Virtues and the Creation of Prosperity*. New York, NY: Free Press.

Hardin, Russell. 1991. "Trusting Persons, Trusting Institutions." In *The Strategy of Choice*, ed. Richard J. Zeckhauser. Cambridge, MA: MIT Press

--1982. *Collective Action*. Baltimore: Resources for the Future by the Johns Hopkins University

Press.

Harford, Thomas, and Leonard Solomon. 1967. " 'Reformed Sinner' and 'Lapsed Saint' Strategies in the Prisoner's Dilemma Game." *Journal of Conflict Resolution* 11:104-109.

Einhorn, Hillel. J., and Hogarth, Robin M. 1985. "Ambiguity and Uncertainty in Probabilistic Inference." *Psychological Review* 92:S225-S250.

Jackman, Robert W., and Ross A. Miller. 1998. "Social Capital and Politics" *Annual Review of Political Science* 1:47-73.

Kelley, Harold H., and Anthony J. Stahelski. 1970. "Social Interaction Basis of Cooperators and Competitors' Beliefs about Others." *Journal of Personality and Social Psychology* 21:190-197.

Knox, R.E., and R.L. Douglas. 1971. "Trivial Incentives, Marginal Comprehension, and Dubious Generalizations from Prisoner's Dilemma Studies." *Journal of Personality and Social Psychology* 12:160-165.

Komorita, Samuel S., and Craig D. Parks. 1994. *Social Dilemmas*. Madison, WI: Brown and Benchmark.

Komorita, Samuel S., J.A. Hilty, and Craig D. Parks. 1991. "Reciprocity and Cooperation in Social Dilemmas." *Journal of Conflict Resolution* 35:494-518.

Komorita, Samuel .S., Craig D. Parks , and L.G. Hulbert. 1992. "Reciprocity and the Induction of Cooperation in Social Dilemmas." *Journal of Personality and Social Psychology* 62:607-17.

Kreps, David. 1990. "Corporate Culture and Economic Theory." In *Perspectives on Positive Political Economy*, ed. James E. Alt and Kenneth A. Shepsle. New York, NY: Cambridge University Press.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108:80-498.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press.

Levi, Margaret. 1988. *Of Rule and Revenue*. Berkeley: University of California Press.

Mete, Mihriye. 1999. "Bureaucratic Behavior in Strategic Environments: Politicians, Taxpayers, and the IRS." Presented at the 1999 Annual Meeting of the Midwest Political Science Association. Chicago, IL.

Miller, Gary. 1992. *Managerial Dilemmas: The Political Economy of Hierarchy.* New York, N.Y.: Cambridge University Press

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. New York, NY: Cambridge University Press.

Orbell, John M., and Robyn M. Dawes. 1991. "A 'Cognitive Miser' Theory of Cooperators' Advantage." *American Political Science Review* 85:515-528.

Orbell, John M., Alphons J.C. van de Kragt, and Robyn M. Dawes. 1988. "Explaining Discussion Induced Cooperation." *Journal of Personality and Social Psychology* 54:811-819.

Oskamp, Stuart. 1971. "Effects of Programmed Strategies on Cooperation in the Prisoner's Dilemma and Other Mixed-Motive Games." *Journal of Conflict Resolution* 15:225-59.

Ostrom, Elinor. 1998. "A Behavioral Approach to the Rational Choice Theory of Collective Action." *American Political Science Review* 92:1-22.

Ostrom, Elinor, Roy Gardner, and James Walker. 1992. "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review* 86:404-17.

Ostrom, Elinor, James Walker, and Roy Gardner. 1994. *Rules, Games and Common-Pool Resources*. Ann Arbor: The University of Michigan Press.

Patchen, Martin. 1987. "Strategies for Eliciting Cooperation from an Adversary." *Journal of Conflict Resolution* 31:164-185.

Payne, John. W., Bettman, James. R., and Johnson, Eric. J. 1993. *The Adaptive Decision Maker*. New York, NY: Cambridge University Press.

Pruitt, Dean G. 1968. "Reciprocity and Credit Building in a Laboratory Dyad." *Journal of Personality* *and Social Psychology* 8:143-47.

Putnam, Robert. 1995. "Bowling Alone: America's Declining Social Capital." *Journal of Democracy* 6: 65-78.

--1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Scholz, John T. and Wayne B. Gray. 1990. "OSHA Enforcement and Workplace Injuries: A Behavioral Approach to Risk Assessment." *Journal of Risk and Uncertainty* 3:283-305.

Scholz, John T. and Mark Lubell. 1998a. "Adaptive Political Attitudes: Duty, Trust, and Fear as Monitors of Tax Policy." *American Journal of Political Science* 42:903-920.

Scholz, John T. and Mark Lubell. 1998b. "Trust and Taxpaying: Testing the Heuristic Approach to Collective Action." *American Journal of Political Science* 42:903-920.

Scholz, John T. and Feng Heng Wei. 1986. "Regulatory Enforcement in a Federalist System." *The American Political Science Review* 80:1249-1270.

Scholz, John T. and B. Dan Wood. 1998. "Controlling the IRS: Principals, Principles, and Public Administration." *American Journal of Political Science* 42:141-162.

Sniderman, Paul M. 1993. "The New Look in Public Opinion Research." In *Political Science: The State of the Discipline II*, ed Ada Finifter. Washington, DC: American Political Science Association.

Taylor, Michael. 1987. *The Possibility of Cooperation.* New York, NY: Cambridge University Press.

Tenbrunsel, Ann E. and David M. Messick. 1999. "Sanctioning Systems, Decision Frames, and Cooperation," *Administrative Science Quarterly* 44:684-707.

Trivers, Robert. 1971. "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology* 46: 35-47.

Wilson, W. 1973. "Reciprocation and Other Techniques for Inducing Cooperation in the Prisoner's Dilemma." *Journal of Conflict Resolution* 15:167-196.