BUILDING RATIONAL COOPERATION*

James Andreoni and Larry Samuelson
Andreoni@wisc.edu and LarrySam@ssc.wisc.edu
Department of Economics
University of Wisconsin
1180 Observatory Drive
Madison, Wisconsin 53706-1321

March 3, 2003

Abstract: We examine a model in which players sometimes prefer to cooperate in the prisoners' dilemma, provided their opponents are sufficiently likely to cooperate. We report the results of an experiment investigating the predictions of this model for equilibrium behavior in a twice-played prisoners' dilemma. We are especially interested in results suggesting that cooperation may be most effectively fostered by appropriately distributing the monetary stakes across the two periods of the interaction.

Contents

*We thank Ted Bergstrom and Bill Sandholm for helpful comments, Emily Blanchard, Tim Classen and Menesh Patel for research assistance, and the Russell Sage Foundation and the National Science Foundation for financial support.

BUILDING RATIONAL COOPERATION

by Jim Andreoni and Larry Samuelson

1 Introduction

Rational cooperation. A fundamental assumption of economic theory is that rational people will not choose strictly dominated strategies. However, experiments consistently find:[1]

- A significant proportion of players cooperate in the one-shot prisoners' dilemma.
- Players are heterogeneous, including some who cooperate under any circumstances, some who defect under any circumstances, and some who appear to be "conditional cooperators," being willing to cooperate if there is sufficient chance that their opponent will do likewise.
- The incidence of cooperation falls over the course of a finitely-repeated prisoners' dilemma, but does not fall to zero.

To say a person is rational is to say that we can specify preferences for which the person's choices consistently coincide with the most-preferred available alternative. The experimental evidence makes it clear that we cannot do so for the prisoners' dilemma if we retain the common assumption that people care only about their own monetary payoffs, so that cooperation is a strictly dominated strategy. One response to such seemingly irrational behavior is to argue that this is a situation beyond the purview of economics. But it is typically more useful to retain the unifying principle of economics—that people can be usefully modeled as making rational choices—while seeking a better understanding of how the observed behavior might be rational. In the case of the prisoners' dilemma, there is no conceptual reason to assume that people care only about their own monetary payoffs. The first step in assessing the experimental evidence is then to ask whether we can rationalize the observed behavior by expanding our notion of preferences to include additional considerations.

The danger here is that the unbridled freedom to tailor preferences to particular circumstances allows us to rationalize virtually any behavior. How do we know when we have uncovered a robust feature of behavior and when

[1] See Andreoni and Miller [2] and the citations within. See Rabin [15] for a discussion of psychological evidence.

we have fortuitously constructed exotic preferences that happen to match some experimental observations?

Our confidence in a specification of preferences invoked to explain a particular behavior is enhanced to the extent that the model makes additional predictions that can be tested experimentally. For example, Andreoni and Miller [3], observing that generosity in the standard dictator game can be explained as rational behavior on the part of dictators who have preferences over both their own and the receiver's monetary payoff, exhibit experimental evidence consistent with the predictions of these preferences in more general dictator games with varying exchange rates. These results encourage further work with such preferences. Andreoni, Castillo and Petrie [1], Binmore, McCarthy, Ponti, Samuelson and Shaked [5] and Falk, Fehr and Fischbacher [11] show that important aspects of play in finite-horizon alternating-offers bargaining games cannot be explained by preferences involving only the relative payoff considerations that were initially introduced to explain seemingly irrational rejections in such games. This suggests that if such rejections are to be explained as optimizing behavior, alternative preference specifications should be investigated.

This paper thus proceeds in three steps. Section 2 constructs a model of individual preferences that yields rational cooperation in the prisoners' dilemma. Section 3 identifies the model's predictions. Sections 4 and 5 report the results of an experiment examining these predictions. Section 6 outlines directions for further work.

This paper. We study a model in which (*i*) players prefer that their opponents cooperate in the prisoners' dilemma, (*ii*) players sometimes prefer to cooperate themselves, (*iii*) players are more likely to cooperate when their opponent is more likely to cooperate, and (*iv*) players differ in the strength of this taste for cooperation. Such preferences are easily formulated to accommodate the experimental observations noted above.

To carry the investigation of these preferences further, we consider two-period games whose stage games are the prisoners' dilemmas shown in Figure 1. Let

$$\lambda = \frac{x_2}{x_1 + x_2}.$$

We consider a class of such twice-played prisoners' dilemma games in which $x_1 + x_2$ is fixed, but λ ranges from zero to one. When $\lambda = 0$, all of the payoffs are concentrated in the first of the two prisoners' dilemmas. As λ increases, the second period becomes relatively more important, with $\lambda = \frac{1}{2}$ corresponding to equal payoffs in the two periods and $\lambda = 1$ corresponding

	C	D
C	$3x_1, 3x_1$	$0, 4x_1$
D	$4x_1, 0$	x_1, x_1

Period 1

	C	D
C	$3x_2, 3x_2$	$0, 4x_2$
D	$4x_2, 0$	x_2, x_2

Period 2

Figure 1: *Stage games for the twice-played prisoners' dilemma, where* $x_1, x_2 \geq 0$.

to all payoffs being concentrated in the second period. With the help of some technical assumptions, designed primarily to ensure that there is sufficient heterogeneity in players' preferences, the model predicts:

- Cooperation will be more prevalent in the first than in the second period of play.
- First-period play for $\lambda = 0$ will match second-period play for $\lambda = 1$.
- The incidence of first-period cooperation increases as λ does.
- Certain outcomes of the game (identified below) become more likely, and others less likely, as λ grows. As a result, the expected monetary payoff from the two-period game initially increases in λ, achieves an interior maximum at a value of λ between zero and one, and then decreases.

Cooperation in the first period, by enhancing an opponent's estimate of one's unobserved taste for cooperation, leads to more opponent cooperation in the second period. This enhances the value of first-period cooperation. As a result, our model shares the common prediction that players are more likely to cooperate at the beginning than at the end of a sequence of prisoners' dilemmas. Our model becomes more interesting when we consider the effects of varying the relative payoffs between the two periods. First, one of the two periods is trivial whenever $\lambda = 0$ or $\lambda = 1$, suggesting that we should observe identical behavior and payoffs from the nontrivial period in each case. More importantly, second-period cooperation is more valuable the higher is λ. As a result, higher values of λ induce agents to cooperate more in the first period as an investment in second-period cooperation. Finally, as λ increases, we trade off increased first-period cooperation for decreased first-period payoffs, as payoffs are shifted to the second period. The combined

effect makes specific predictions concerning the path of play and causes monetary payoffs to be minimized when $\lambda = 0$ or $\lambda = 1$, and to achieve an interior maximum.

Building cooperation. We view this work as a small step toward an understanding of how institutions might be designed to facilitate economic and social interactions.

It is well recognized that the performance of an economy can depend importantly upon the institutions within which interactions take place. Much of the literature has concentrated on the more formal of these institutions, including laws creating property rights and structuring economic activity, legal systems to enforce these laws, financial systems that facilitate trade, educational systems that produce information, and so on.

Our work is motivated by the observation that the functioning of a modern economy depends upon the consistent willingness to forgo opportunistic behavior in favor of cooperation. As observed by Arrow [4], virtually every economic exchange requires someone to forsake individual advantage, even something so mundane as buying a loaf a bread—there is invariably a moment at which one party has the loaf of bread as well as the money, and has an opportunity to take both.

One response to the potential for such opportunistic behavior is to write contracts or pass laws to deter opportunism with legally enforced sanctions, i.e., to rely on formal institutions. But these remedies are costly and clumsy. The more a society can rely on voluntary cooperation to sustain economic interactions the richer the society will be. Indeed, social researchers have recently argued that the ability to harness voluntary cooperation, often called "social capital" (Putnam [14]), is sometimes more important than physical or human capital in building modern economies (Knack and Keefer [13], Whiteley [18]).

We believe that cooperation can be encouraged or attenuated by the environment surrounding an interaction, much of which lies beyond formal economic institutions. But what organizational arrangements encourage people to cooperate with one another? How do we design social and economic institutions to build and sustain cooperation? Understanding the preferences that induce people to cooperate is the first step toward answering these questions.

The work reported in this paper is especially relevant to the proposition that cooperation is fostered in relationships whose stakes are appropriately distributed across periods. An associate who accomplishes one task well

may be given more discretion and responsibility. Wary countries begin with cultural exchanges, work up to economic ties, and then negotiate military treaties. Starting a relationship with small stakes allows people an opportunity to exhibit a willingness to cooperate, and to assess the propensity of others to cooperate, when the risk of doing so is mitigated by relatively small payoffs. These investments can pay off in the form of mutual cooperation for subsequent, larger stakes.

While it seems intuitive that "starting small" can foster cooperation, the argument involves some subtleties that must be treated with care. The same low payoffs that mitigate the risk of building cooperation also make it more attractive to cooperate now in order to tempt others into cooperation so that they can be exploited at higher stakes, prompting suspicion of initial cooperators. Identifying circumstances conducive to cooperation thus requires precise theoretical modeling and experimental work.

Relationship to the literature. The literature contains two models of relationships that start small that are relevant to our work.[2] First, consider an infinitely-repeated prisoners' dilemma. Equilibria featuring cooperation exist if discount factors are sufficiently high. Causing the stakes of the game to increase over time increases the effective discount factor. It may then be that discount factors are too low to sustain cooperation if the stakes of the game remain unchanged throughout the horizon, but cooperation can be sustained if the stakes increase.[3]

More closely related ideas appear in Watson's [16, 17] study of the infinitely repeated prisoners' dilemma. In his simplest model, there are L and H agents, with the former having discount factors that make them more inclined to defect. In equilibrium, L agents defect immediately, while the stakes of the game gradually increase to a steady-state level. The rate at which the stakes increase is fixed by an incentive-compatibility constraint that the L agents not prefer to delay their defection to a period with larger

[2]Binmore, Proulx, Samuelson and Swierzbinski [6] present experimental results in which players are more likely to trust a randomly chosen opponent if they must first risk relatively small amounts to do so, building up to risking larger amounts, than if the high-stakes trust opportunities come first.

[3]Datta [9] examines a version of the repeated prisoners' dilemma in which players can abandon their existing partners in favor of new ones. Punishments for defection are ineffective, since one can always seek a new partner with whom to resume cooperation. However, if the stakes of the game increase over time, then it is costly to abandon current high-stakes cooperation in order to start over at low stakes, restoring the possibility of cooperation. Similar ideas appear in Carmichael and MacLeod [8] and Ghosh and Ray [12].

payoffs. If the path of increasing stakes is to be designed to maximize H-player payoffs, a trade-off appears. The smaller are initial payoffs, the less costly are the initial L defections, but the longer must H agents wait until achieving cooperation at large stakes.[4] This is similar to our result that shifting payoffs to the second period increases first-period cooperation, but at smaller stakes.

2 The Model

2.1 Preferences

Our analysis begins with the assumption that, given a specification of the monetary payoffs for a one-shot prisoners' dilemma, an agent's utility from cooperating (C) and defecting (D) is given by

$$C: \qquad \pi(C, \rho, \alpha) + \theta_C \tag{1}$$

$$D: \qquad \pi(D, \rho, \alpha) + \theta_D. \tag{2}$$

where

$$\pi(z, \rho, \alpha) : \{C, D\} \times [\underline{\alpha}, \overline{\alpha}] \times [\underline{\alpha}, \overline{\alpha}] \rightarrow I\!R \tag{3}$$

identifies an agent's expected utility as a function of the action $z \in \{C, D\}$ chosen by the agent, the probability ρ with which the agent's opponent cooperates,[5] and the value α of a parameter characterizing the agent that we interpret shortly.

The function $\pi(z, \rho, \alpha)$ is an expected utility in two senses. First, the realized utility depends upon the opponents' action, which we incorporate by writing the expected utility as a function of the probability ρ that the opponent cooperates. Second, we assume that the utility of cooperating (or defecting) is perturbed by a random variable Θ_C (or Θ_D). The random variables Θ_C and Θ_D are independent and have zero means, with distrubutions that are differentiable and strictly increasing on the reals, so that the random variables exhibit full support and no mass points.

The realized values θ_C and θ_D of these random variables are drawn before the agent makes his choice. In contrast, the agent's choice must be made without knowing the opponent's action. As a result, the utilities in (1)–(2)

[4] Similar ideas appear in Blonski and Probst [7] and Diamond [10].

[5] Though we interpret ρ as the probability with which the opponent cooperates, it simplifies the presentation and notation to define $\pi(z, \rho, \alpha)$ for values of ρ lying in $[\underline{\alpha}, \overline{\alpha}]$, where $\underline{\alpha} < 0 < 1 < \overline{\alpha}$.

are defined as a function of the realizations of θ_C and θ_D but remain an expectation over the opponent's action.

The random variables Θ_C and Θ_D, reflecting a realization that the function $\pi(z, \rho, \alpha)$ may not capture every detail of agents' preferences, will be important when interpreting the experimental results. However, our working assumption is that $\pi(z, \rho, \alpha)$ provides a useful approximation of preferences. We thus focus on results that hold when Θ_C and Θ_D are *sufficiently small*, meaning that the distributions of Θ_C and Θ_D are sufficiently close (in the topology of weak convergence) to degenerate distributions that put unitary mass on zero.

The utility function $\pi(z, \rho, \alpha)$ and the distributions of the random variables Θ_C and Θ_D will depend upon the monetary payoffs of the prisoners' dilemma under consideration. We assume that $\pi(C, \rho, \alpha)$ and $\pi(D, \rho, \alpha)$ are homogeneous of degree one in monetary payoffs and that the random variables Θ_C and Θ_D are similarly homogeneous.[6] Equilibrium play in a one-shot prisoners' dilemma would thus be unaffected by the stakes of the game, in the sense that multiplying monetary payoffs by a common factor would leave an equilibrium unaffected.[7] We accordingly suppress notation for the monetary payoffs of the game.

We assume that a player's willingness to cooperate depends upon the behavior of the player's opponent:

Assumption 1 *For all* (z, ρ, α)*,*

$$\frac{d\pi(z, \rho, \alpha)}{d\rho} > 0 \tag{4}$$

and

$$\frac{d[\pi(C, \rho, \alpha) - \pi(D, \rho, \alpha)]}{d\rho} > 0. \tag{5}$$

A player thus prefers that his opponent cooperate, and finds cooperation relatively more attractive the more likely is the opponent to cooperate. Figure 2 illustrates preferences that are consistent with this assumption. If condition (5) fails, then defection becomes more attractive the more likely

[6] Hence, if the random variable Θ'_C pertains to a prisoners' dilemma whose monetary payoffs are k (> 0) times those for the prisoners' dilemma corresponding to Θ_C, then $\Theta'_C(\omega) = k\Theta_C(\omega)$, where these random variables are defined on a common state space Ω with $\omega \in \Omega$. As a result, $\text{prob}\{\theta_C \in [\overline{\theta}, \underline{\theta}]\} = \text{prob}\{\theta'_C \in [k\overline{\theta}, k\underline{\theta}]\}$.

[7] Our experiment involves only nonnegative monetary payoffs, weakening this assumption somewhat by obviating the need to compare gains and losses.

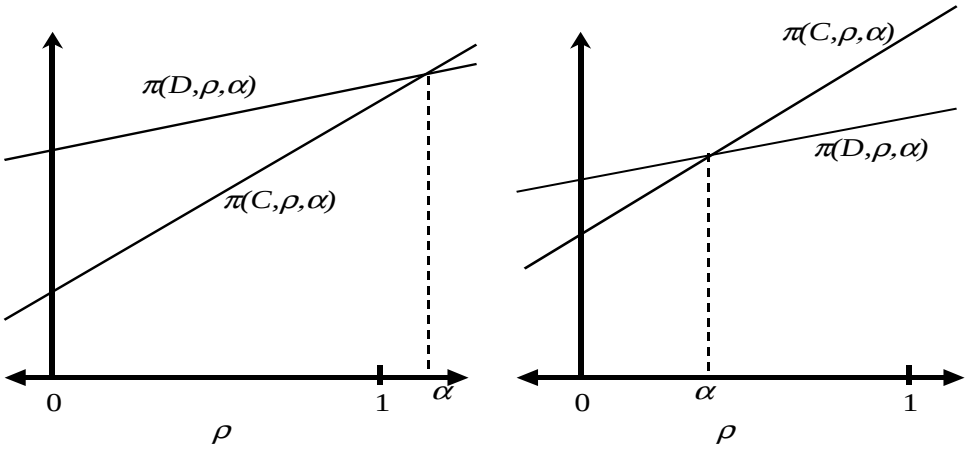

Figure 2: *Possible utilities of cooperation* ($\pi(C, \rho, \alpha)$) *and defection* ($\pi(D, \rho, \alpha)$) *as a function of probability* ρ *that an opponent cooperates (cf. note 5).*

is an opponent to cooperate, in which case the agent not only delights in fleecing others, but also delights in being fleeced.

We assume that the labels represented by α are assigned to players so that a player characterized by α is indifferent between C and D, given $\theta_C = \theta_D = 0$, when his opponent plays C with probability α, i.e.,

$$\pi(C, \alpha, \alpha) = \pi(D, \alpha, \alpha).$$

Equivalently, α is the probability of opponent cooperation above which a player of characteristic α prefers to cooperate rather than defect. Despite calling α a probability, we do not restrict α to lie within $[0, 1]$ (cf. Figure 2 and note 5). Instead, a value of $\alpha < 0$ denotes an agent for whom C is a dominant strategy in the one-shot prisoners' dilemma (given $\theta_C = \theta_D = 0$). A value of $\alpha > 1$ denotes an agent for whom D is a dominant strategy in the one-shot prisoners' dilemma (again, absent perturbations), as in the left panel of Figure 2. A value $\alpha \in (0, 1)$ is an agent who sometimes prefers C and sometimes D, depending upon the probability that the opponent cooperates, as in the right panel of Figure 2. We refer to the latter agents as *conditional cooperators*.

The possibility that players may have different preferences will be captured by allowing players to be characterized by different values of α. We find it convenient to refer to a player characterized by value α as "player

α." We think of α as being a characteristic of a player that is fixed, while the perturbations θ_C and θ_D are drawn anew each time the game is played.

We assume:

Assumption 2 *If $\rho - \alpha = \rho' - \alpha'$, then*

$$\pi(C, \rho, \alpha) - \pi(D, \rho, \alpha) = \pi(C, \rho', \alpha') - \pi(D, \rho', \alpha').$$

This assumption ensures that the parameter α captures all of the information we need about the differing preferences of different agents. Players with different values of α are characterized by utility differences $\pi(C, \rho, \alpha) - \pi(D, \rho, \alpha)$ (as a function of ρ) that are horizontal shifts of one another.

We assume that each player's value of α is drawn independently according to the distribution function $F : [\underline{\alpha}, \overline{\alpha}] \to [0, 1]$:

Assumption 3 *The distribution function $F(\alpha)$ is differentiable on $[\underline{\alpha}, \overline{\alpha}]$, with $\underline{\alpha} < 0 < 1 < \overline{\alpha}$ and*

$$0 < \frac{dF(\alpha)}{d\alpha} < 1. \tag{6}$$

The differentiability of F, by ruling out mass points of agents of a single type, plays a key role in ensuring that we have equilibria in pure strategies. The assumption that F has a slope less than one ensures uniqueness of equilibrium in the one-shot game.

We let f denote the density of F. The assumption that F is strictly increasing on $[\underline{\alpha}, \overline{\alpha}]$ ensures that

$$\begin{aligned} F(0) &> 0 \\ 1 - F(1) &> 0, \end{aligned}$$

where $F(0)$ is the proportion of "committed cooperators," who prefer cooperation regardless of their opponent's action (given $\theta_C = \theta_D = 0$), $1 - F(1)$ is the proportion of "committed defectors" who prefer defection regardless of their opponent's action, and $F(1) - F(0)$ is the proportion of conditional cooperators. Notice that among the committed cooperators and defectors, there is a sense in which those with more extreme values of α (lower in the case of cooperation, higher in the case of defection) are "more committed." This is again useful in avoiding mixed strategies.

All of the assumptions we make on F in this paper are satisfied if F is a uniform distribution on $[\underline{\alpha}, \overline{\alpha}]$.

2.2 Equilibrium of the One-Shot Game

We assume that players matched to play the game know their own preferences, including their values of α and the realized values of θ_C and θ_D, but know only that their opponent's values of α, θ_C and θ_D are independently drawn from the corresponding distributions. The appropriate equilibrium concept in the one-shot prisoners' dilemma is Bayesian-Nash equilibrium.

Let $\delta = \theta_C - \theta_D$. Then δ is the realization of a random variable whose distribution converges weakly to a unitary mass on zero as do the distributions of Θ_C and Θ_D. Behavior will depend only upon δ rather than upon the values of θ_C and θ_D, and we will refer to a player as being characterized by a value of α and a realized value of δ. We have:[8]

Proposition 1 *Let Assumptions 1–3 hold. Then:*

(1.1) There exists a unique Bayesian-Nash equilibrium in the one-shot game, characterized by an increasing function $\Delta(\alpha)$ such that a player characterized by (δ, α) cooperates if $\delta > \Delta(\alpha)$ and defects if $\delta < \Delta(\alpha)$.

(1.2) In the limit in which Θ_C and Θ_D are zero, the equilibrium is characterized by a value $\alpha^* \in (0,1)$ satisfying $F(\alpha^*) = \alpha^*$, with player α cooperating if $\alpha < \alpha^*$ and defecting if $\alpha > \alpha^*$.

Hence, players with smaller values of α and larger values of δ are more likely to cooperate.

Proof. It is immediate from (5) and Assumption 2 that the equilibrium must be characterized by an increasing function $\Delta(\alpha)$ such that a player characterized by (δ, α) cooperates if $\delta > \Delta(\alpha)$ and defects if $\delta < \Delta(\alpha)$.

To establish existence, note that any probability $\rho \in [0,1]$ that a randomly drawn opponent cooperates induces a unique such function that we can write as $\tilde{\Delta}(\alpha, \rho)$, defined by

$$\pi(C, \rho, \alpha) - \pi(D, \rho, \alpha) + \tilde{\Delta}(\alpha, \rho) = 0,$$

and that such a function induces a probability of cooperation given by

$$R(\rho) = \text{prob}\{(\delta, \alpha) : \delta > \tilde{\Delta}(\alpha, \rho)\}.$$

An equilibrium exists if we can find a value of ρ such that

$$\rho = R(\rho).$$

[8] Throughout, we characterize equilibria only up to measure-zero sets of agents who are indifferent.

Existence then follows from the observation that $R(\rho)$ is increasing and continuous, with $R(0) > 0$ and $R(1) < 1$, ensuring that $\rho - R(\rho) = 0$ has at least one solution within the unit interval.

To establish uniqueness, note that

$$R(\rho) = \int_{-\infty}^{\infty} \int_{\underline{\alpha}}^{h(\delta,\rho)} f(\alpha) d\alpha \; g(\delta) d\delta,$$

where g is the density of δ and $h(\delta, \rho)$ identifies that value of α for which an agent with preference shock δ is indifferent between cooperating and defecting given probability ρ of opponent cooperation, or

$$\pi(C, \rho, h(\delta, \rho)) = \pi(D, \rho, h(\delta, \rho)).$$

Then

$$\frac{dR(\rho)}{d\rho} = \int_{-\infty}^{\infty} f(h(\delta, \rho)) \frac{dh(\delta, \rho)}{d\rho} g(\delta) d\delta.$$

Assumption (2) implies that $dh(\delta, \rho)/d\rho = 1$. Assumption 3 ensures that $f(h(\delta, \rho)) < 1$, which gives $dR(\rho)/d\rho < 1$. This ensures that $\rho - R(\rho)$ has a unique fixed point, and hence that there is a unique equilibrium.

Now consider the limiting case in which Θ_C and Θ_D are zero. Then the equilibrium is characterized by a value α^* such that larger values of α defect and smaller ones cooperate, with player α^* being indifferent between C and D. Given the differentiability of F, the equilibrium proportion of cooperators will then be $F(\alpha^*)$. The indifference of α^* requires $F(\alpha^*) = \alpha^*$. The existence of α^* is straightforward, while the fact that $dF(\alpha)/d\alpha < 1$ ensures that there is a unique such α^*. ||

Figure 3 illustrates the equilibrium for the limiting case in which the perturbations Θ_C and Θ_D are arbitrarily small. The equilibrium value α^* clearly depends on the distribution F of players' indifference points, which reflects the characteristics of the players and the specification of the monetary payoffs of the prisoners' dilemma.

Consider two populations characterized by distributions F and F' over α with F first-order stochastically dominating F', so players in population F are characterized by higher values of α and hence are less likely to cooperate. Then for any fixed $\tilde{\Delta}(\alpha, \rho)$, F will induce a smaller value of $R(\rho)$ than does F'. This gives:

Corollary 1 *If distribution F (over α) first-order stochastically dominates F', then populations characterized by these two distributions induce equilibria with $\Delta(\alpha) > \Delta'(\alpha)$, and hence F' induces more cooperation than does F.*

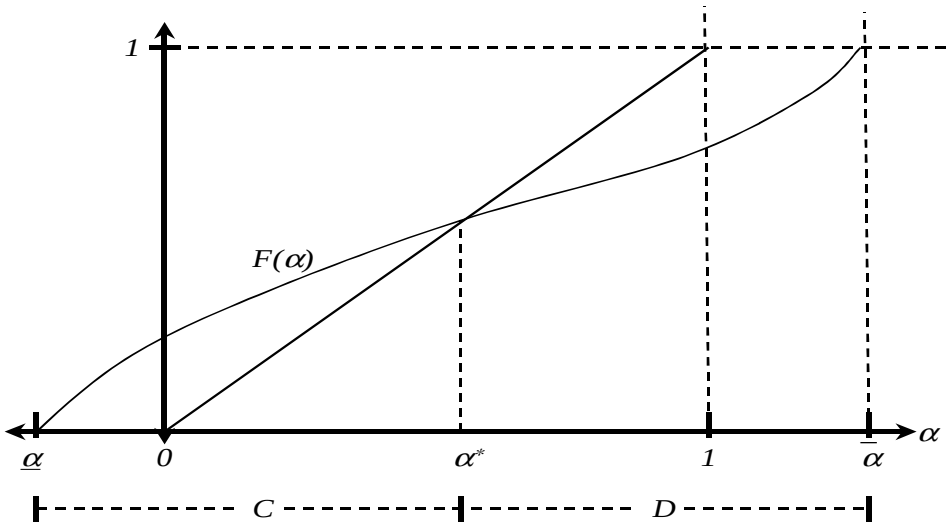

Figure 3: *Equilibrium of one-shot prisoners' dilemma when* Θ_C *and* Θ_D *are zero.*

2.3 Equilibrium of the Twice-Played Game.

We now consider the perfect Bayesian equilibria of the twice-played prisoners' dilemma. We denote the equilibrium of the single-stage game by $\Delta^*(\alpha)$, or by simply α^* when considering the noiseless limit. Unless otherwise noted, we restrict attention to values of $\lambda \in (0, 1)$, so that both periods have nontrivial payoff implications.

There are potentially multiple equilibria of the two-stage game, some of which can be counterintuitive:

Example. Let $\lambda = \frac{1}{2}$ and hence $x_1 = x_2$, so that first-period and second-period payoffs are identical, and (to simplify the example) let Θ_C and Θ_D each place unitary mass on a value of zero. We show that there are circumstances under which it is a perfect Bayesian equilibrium for all agents to play the following strategy for the two-period game:

1. Defect in the first period.
2. If first-period play yields (D, D), then play the equilibrium of Proposition 1 in the second period.

3. If an opponent plays C in the first period, draw the inference that the opponent's value of α exceeds one (and hence that the opponent will defect in the next period). If the opponent cooperates in the first period, a player thus chooses C in the second period if and only if $\alpha < 0$. The first-period (out-of-equilibrium) cooperating player, anticipating this response, cooperates in the second period if and only if $\alpha < F(0)$.

Given that all agents play D in the first period, (D, D) outcomes are uninformative, making the continuation strategy of playing the one-shot equilibrium prescribed by Proposition 1 optimal in the second period. Similarly, given that equilibrium first-period play calls for defection, a perfect Bayesian equilibrium allows the inference that the opponent's value of α exceeds one if the opponent cooperates in the first period. Since a player for whom $\alpha > 1$ finds defection a dominant strategy in a one-shot game, the best response to such an inference is to cooperate if and only if one's own value of α is less than zero, and hence the second-period behavior conditional on a first-period choice of C by either player is again optimal.

It then remains only to verify that defection is optimal in the first period, given the prescribed continuation behavior. Notice first that D leads to a higher probability that one's opponent cooperates in the second period. Hence, for any agent for whom $\alpha > 0$, defection is a strict best response in the first period and produces a higher continuation value than does C, making D optimal in the first period. First-period defection is optimal for agents with $\alpha < 0$ if

$$\pi(C, 0, \alpha) - \pi(D, 0, \alpha) \leq \pi(C, \alpha^*, \alpha) - \pi(C, F(0), \alpha),$$

where the left side is the first-period gain from switching to cooperation and the right side is the second-period sacrifice from doing so. We can easily find specifications of the problem for which this inequality holds. For example, let F be uniformly distributed on $[-\frac{1}{4}, \frac{5}{4}]$ and let

$$\begin{aligned} \pi(C, \rho, \alpha) &= \rho - \alpha \\ \pi(D, \rho, \alpha) &= \tfrac{1}{2}(\rho - \alpha). \end{aligned}$$

Then $\alpha^* = \frac{1}{2}$ and $F(0) = \frac{1}{6}$, and the required inequality is, for $\alpha < 0$,

$$(-\alpha) - (-\tfrac{1}{2}\alpha) \leq (\tfrac{1}{2} - \alpha) - (\tfrac{1}{6} - \alpha)$$

or $-\frac{1}{2}\alpha < \frac{1}{3}$, which holds for all $\alpha \in [-\frac{1}{4}, 0]$. The posited strategies are thus a perfect Bayesian equilibrium. ||

This equilibrium has the property that cooperation in the first period leads players to believe that cooperation is *less* likely in the second period. We regard such equilibria as counterintuitive and accordingly restrict attention to monotonic equilibria:

Definition 1 *An equilibrium of the twice-played prisoners' dilemma is* monotonic *if the probability that player i's opponent cooperates in period two is at least as high when player i cooperates in period one as when player i defects.*

We have a convenient characterization of first-period behavior in monotonic equilibria:

Lemma 1 *In a monotonic equilibrium, there is an increasing function* $\Delta_1(\alpha)$ *such that in the first period, an agent characterized by* (δ, α) *cooperates if and only if* $\delta > \Delta_1(\alpha)$.

Proof. Consider an agent α and value δ that prompts cooperation in the first period. Let ρ_1 be the probability of opponent cooperation in the first period. Let $V(z, \alpha)$ be the expected value of the second period of the game (conditional on the equilibrium) to an agent of type α who takes action $z \in \{C, D\}$ in the first period, where the expectation is taken over the likely type (and hence actions) of the opponent. Then the optimality of player α's choice requires

$$\pi(C, \rho_1, \alpha) + V(C, \alpha) + \delta \geq \pi(D, \rho_1, \alpha) + V(D, \alpha). \tag{7}$$

Letting $\Delta_1(\alpha)$ be the value of δ that satisfies this relationship with equality, player α will cooperate in period one iff $\delta > \Delta_1(\alpha)$. Now let $\alpha' < \alpha$. Assumption 2 implies that, for a monotonic equilibrium,

$$V(C, \alpha) - V(D, \alpha) \leq V(C, \alpha') - V(D, \alpha').$$

Using this inequality and Assumption 2 again, (7) implies

$$\pi(C, \rho_1, \alpha') + V(C, \alpha') + \delta \geq \pi(D, \rho_1, \alpha') + V(D, \alpha'),$$

ensuring that agent α' cooperates for any value of δ that prompts α to cooperate, and hence that $\Delta_1(\alpha)$ is increasing. ||

We next show that players are more inclined to cooperate in the first period of a two-period game than in a one-shot game:[9]

[9]Players will be *strictly* more inclined to cooperate in the first period of a two-period game than in the one-shot game (i.e., $\Delta_1(\alpha) < \Delta^*(\alpha)$) if the equilibrium of the former is strictly monotonic, meaning that first-period cooperation strictly increases the incidence of second-period cooperation.

Lemma 2 *In a monotonic equilibrium,*

$$\Delta_1(\alpha) \leq \Delta^*(\alpha).$$

Proof. These functions satisfy:

$$\begin{aligned} \Delta^*(\alpha) &= \pi(D, \rho^*, \alpha) - \pi(C, \rho^*, \alpha) && (8) \\ \Delta_1(\alpha) &= \pi(D, \rho_1, \alpha) + V(D, \alpha) - \pi(C, \rho_1, \alpha) - V(C, \alpha), && (9) \end{aligned}$$

where ρ^* and ρ_1 are the equilibrium first-period probabilities that the opponent cooperates. Define $\tilde{\Delta}_1(\alpha, \rho)$ to satisfy

$$\pi(C, \rho, \alpha) - \pi(D, \rho, \alpha) + V(C, \alpha) - V(D, \alpha) + \tilde{\Delta}_1(\alpha, \rho) = 0,$$

and

$$R_1(\rho) = \text{prob}\{(\delta, \alpha) : \delta > \tilde{\Delta}_1(\alpha, \rho)\}.$$

These are analogous to the functions $\tilde{\Delta}(\alpha, \rho)$ and $R(\rho)$ used in the proof of Proposition 1. Then consider $\rho_1 = \rho^*$. From (8)–(9) and the fact that, in a monotonic equilibrium $V(C, \alpha) \geq V(D, \alpha)$ (since current cooperation weakly increases the likelihood that the opponent cooperates in the next period), we then have $R_1(\rho_1) \geq R(\rho_1) = R(\rho^*) = \rho^* = \rho_1$, and hence $R_1(\rho_1) \geq \rho_1$. Once again, $dR_1(\rho)/d\rho < 1$, and hence the equilibrium value of ρ_1 must satisfy $\rho_1 > \rho^*$. Achieving such a higher incidence of first-period cooperation requires $\Delta_1(\alpha) \leq \Delta^*(\alpha)$. ||

We now turn to the analysis of the equilibrium in the two-period game. As always, the difficulty in establishing the existence of an equilibrium lies in showing that second-period behavior is nicely behaved as a function of first-period strategies. This in turn hinges upon showing that the second period features a unique equilibrium. Our approach is to offer assumptions sufficient to ensure that the second-period equilibrium is unique when Θ_C and Θ_D are zero, and then to verify that such uniqueness continues to hold when Θ_C and Θ_D are sufficiently small.

The assumptions are:

Assumption 4

$$\begin{aligned} F(0) &\leq (\alpha^*)^2 && (10) \\ F(1) &< 2\alpha^* - (\alpha^*)^2 && (11) \\ f(\alpha) &< \frac{F(\alpha)}{\alpha} \qquad \forall \alpha \in [\alpha^*, \overline{\alpha}]. && (12) \end{aligned}$$

The first two conditions require that there be not too many committed cooperators, and that there be sufficiently many committed defectors. Our intuition concerning the twice-played game is based on the presumption that conditional cooperators may cooperate in the first period to encourage cooperation on the part of their opponents and may modify their second-period behavior in response to inferences drawn about their opponent. The first feature becomes unimportant if there are too many committed cooperators, while the second becomes unimportant if there are too few committed defectors. The third condition ensures that too much probability mass cannot become concentrated on a small set of types who are biased toward defection.

Proposition 2 *Let Assumptions 1–4 hold. If* $|d^2F(\alpha)/d\alpha^2|$, Θ_C *and* Θ_D *are sufficiently small and* $\lambda \in (0,1)$*, then:*

(2.1) Let $\Delta_1(\alpha)$ *be increasing with* $\Delta_1(\alpha) < \Delta^*(\alpha)$ *for all* α *and assume that, in the first period, player* (δ, α) *cooperates if and only if* $\delta > \Delta_1(\alpha)$*. Then optimal second-period behavior is uniquely determined and is given by:*

1. *If both players cooperate in the first period, then there exists an increasing function* $\Delta_{\{2,CC\}}(\alpha) < \Delta_1(\alpha)$ *such that a player characterized by* (δ, α) *cooperates in the second period if and only if* $\delta > \Delta_{\{2,CC\}}(\alpha)$*.*

2. *If both players defect in the first period, then there exists an increasing function* $\Delta_{\{2,DD\}}(\alpha) > \Delta_1(\alpha)$ *such that a player characterized by* (δ, α) *cooperates in the second period if and only if* $\delta > \Delta_{\{2,DD\}}(\alpha)$*.*

3. *If player* i *cooperates and* j *defects in the first period, then there exist increasing functions* $\Delta_{\{2,CD\}}(\alpha) > \Delta_1(\alpha)$ *and* $\Delta_{\{2,DC\}}(\alpha) < \Delta_1(\alpha)$ *such that player* i *(*j*) cooperates in period 2 if and only if* $\delta(i) > \Delta_{\{2,CD\}}(\alpha(i))$ *(*$\delta(j) > \Delta_{\{2,DC\}}(\alpha(j))$*).*

(2.2) A monotonic equilibrium of the twice-played game exists.

Remark. If Θ_C and Θ_D are zero, then the first period is characterized by a value $\alpha_1 > \alpha^*$ such that in the first period, player α cooperates if and only if $\alpha < \alpha_1$. Second-period behavior is then given by:

1. Let α_2 be the unique solution to

$$\frac{F(\alpha_2)}{F(\alpha_1)} = \alpha_2 \tag{13}$$

if $\alpha_1 > 1$, and let $\alpha_2 = 1$ if $\alpha_1 < 1$. If both players cooperate in the first period, then player α cooperates in the second period if and only if $\alpha < \alpha_2$.

2. If both players defect in the first period, then player α cooperates in the second period if and only if $\alpha < 0$.

3. If player i cooperates and j defects in the first period, then player i (j), characterized by $\alpha(i)$ ($\alpha(j)$) cooperates in the second period if and only if $\alpha(i) < 0$ ($\alpha(j) < F(0)/F(\alpha_1)$).

Notice that Proposition 2 does not establish the uniqueness of a monotonic equilibrium in the two-period game. Instead, there may be multiple monotonic equilibria corresponding to different values of $\Delta_1(\alpha)$, the first period cooperation cut-off, with a unique equilibrium conditional on each value of $\Delta_1(\alpha)$. Notice, however, that if there are two such equilibrium functions, one of them must exceed the other for every value of α. If there are multiple equilibria, we refer to the equilibrium characterized by the smallest $\Delta_1(\alpha)$ (largest propensity for first-period cooperation) as the *maximally cooperative* equilibrium.

Proof. (2.1) If a player cooperates (defects) in period 1, then beliefs about that player's value of α are given by a full-support distribution on $[\underline{\alpha}, \overline{\alpha}]$ that is first-order stochastically dominated by (first-order stochastically dominates) the prior distribution F. Corollary 1 thus ensures that the second-period equilibrium following a first-period outcome of either CC or DD must take the form described in the proposition. It is also immediate that following an outcome of CD, second period behavior is described by a pair of increasing functions $\Delta_{\{2,CD\}}(\alpha)$ and $\Delta_{\{2,DC\}}(\alpha)$. It remains to establish that the latter bear the claimed relationship to $\Delta_1(\alpha)$ and to establish uniqueness.

We present the argument for the case in which Θ_C and Θ_D are zero, verifying as we proceed that it can be extended to cases in which Θ_C and Θ_D are nonzero but small. We first note that second-period posterior beliefs about an opponent's type (given $\Theta_C = \Theta_D = 0$) are given by

$$\begin{array}{ll} \frac{F(\alpha)}{F(\alpha_1)} & \text{on } [\underline{\alpha}, \alpha_1] \\ 1 & \text{on } [\alpha_1, \overline{\alpha}] \end{array} \quad \text{if } C \text{ observed}$$

$$\begin{array}{ll} 0 & \text{on } [\underline{\alpha}, \alpha_1] \\ \frac{F(\alpha)-F(\alpha_1)}{1-F(\alpha_1)} & \text{on } [\alpha_1, \overline{\alpha}] \end{array} \quad \text{if } D \text{ observed} \qquad (14)$$

Figure 4 illustrates these posteriors. If Θ_C and Θ_D are nonzero, then the posterior beliefs after an observation of C or D in the first period each have full support on $[\underline{\alpha}, \overline{\alpha}]$, converging to those given in (14) as Θ_C and Θ_D get small.

We now verify that the second-period strategies yield an equilibrium. If both agents cooperate in the first period, then the optimality and uniqueness of the second-period equilibrium is established by Proposition 1, with the following modification. The proof of Proposition 1 uses the assumption that $f(1) < 1$ to establish the uniqueness of an equilibrium in the one-shot game. When examining the second period of the two-period game, we need to establish the existence of a unique α_2 for which

$$\frac{F(\alpha_2)}{F(\alpha_1)} = \alpha_2. \tag{15}$$

It suffices for uniqueness that, for any α_1,

$$\frac{F(\alpha)}{F(\alpha_1)} = \alpha \quad \Rightarrow \quad \frac{f(\alpha)}{F(\alpha_1)} \leq 1. \tag{16}$$

This ensures that the function $F(\alpha)/F(\alpha_1)$ can intersect the diagonal at most once. (The bottom panel of Figure 4 shows an example where there is no such intersection. Figure 5 shows a case, requiring $\alpha_1 > 1$, in which there is such an intersection.) Using the first equality in (16) to eliminate α_1 from the second, (16) is implied by (12).

Extending this result to the case in which Θ_C and Θ_D are nonzero requires establishing the uniqueness of the function $\Delta_{\{2,CC\}}(\alpha)$, which in turn requires showing that the second-period counterpart of $R(\rho)$ (see the proof of Proposition 1) has a slope less than one. This clearly holds if Θ_C and Θ_D are small, with the slope of $R(\rho)$ converging to the slope of $F(\alpha)/F(\alpha_1)$ as Θ_C and Θ_D converge to zero.

If both agents defect in the first period, then both expect defection in the second period, making it optimal to cooperate only if $\alpha < 0$. The proof of Proposition 1 can again be mimicked to show that if both players defect in the first period, then the only possible equilibrium continuation play is for each to cooperate if and only if $\alpha < 0$. Again, the extension to small Θ_C and Θ_D follows in a straightforward way from the convergence of posterior beliefs to those given in (14).

If player i cooperates and player j defects in period one, then i expects his opponent to certainly defect in the second period, and hence finds it optimal to cooperate if and only if $\alpha < 0$. Agent j expects cooperation

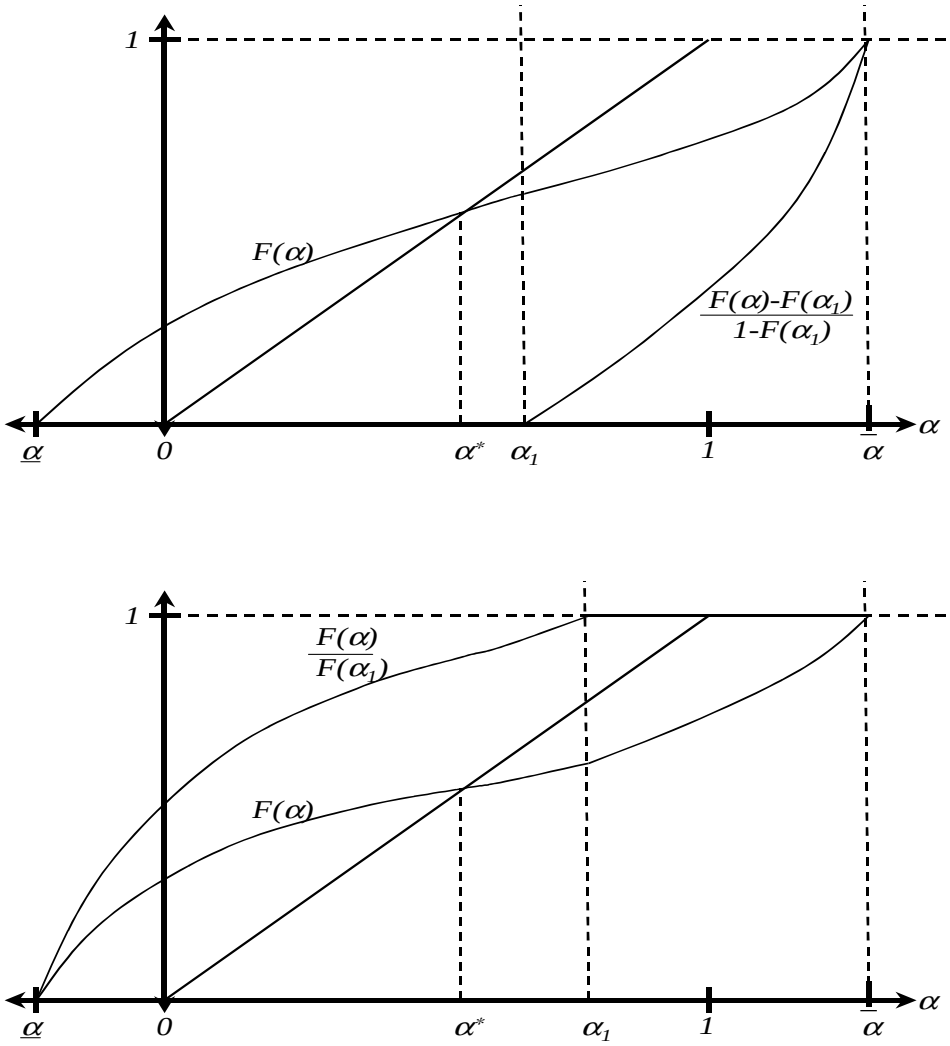

Figure 4: *Posterior beliefs in equilibrium of twice-played prisoners' dilemma, for the case of Θ_C and Θ_D equal to zero, given an observation of D (top) or C (bottom).*

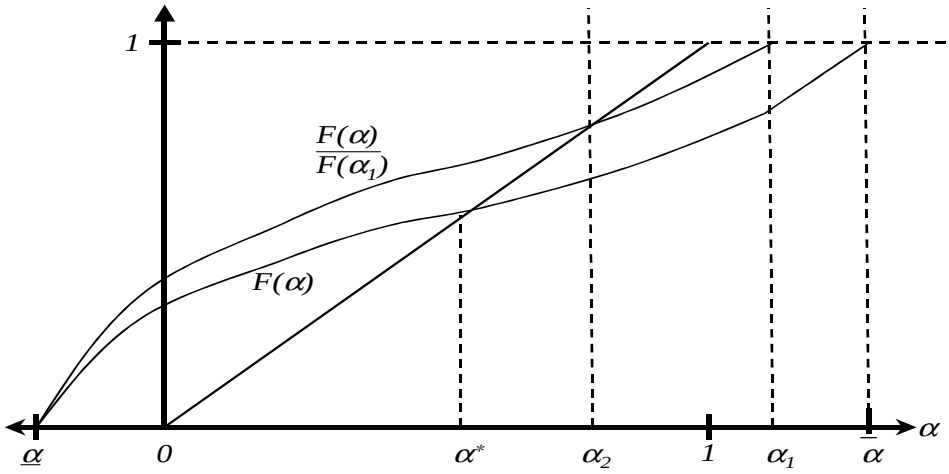

Figure 5: *Possible second-stage equilibrium following* (C, C).

with probability $F(0)/F(\alpha_1)$, making it optimal to defect in the second period for all $\alpha > \alpha_1$ if $F(0)/F(\alpha_1) \leq \alpha_1$. It suffices for this inequality that $F(0)/F(\alpha^*) < \alpha^*$ which, using $F(\alpha^*) = \alpha^*$, is ensured by (10). To address the case of nonzero Θ_C and Θ_D, notice that the functions $\Delta_{\{2,CD\}}(\alpha)$ and $\Delta_{\{2,DC\}}(\alpha)$ are completely determined once one identifies the values $\alpha(CD)$ and $\alpha(DC)$ at which $\Delta_{\{2,CD\}}(\alpha(CD)) = 0 = \Delta_{\{2,DC\}}(\alpha(DC))$. Once again, the convergence of posterior expectations ensures that there exists an equilibrium in which $\alpha(DC)$ converges to zero and $\alpha(DC)$ to $F(0)/F(\alpha_1)$.

The prescribed second-period behavior thus constitutes an equilibrium. Uniqueness is established in the appendix.

(2.2) The existence of an equilibrium in the two-period game now follows from noting that second-period behavior is both unique and a continuous function of the first-period value of α_1, allowing us to apply a fixed-point argument analogous to that used to establish existence in the one-shot game. ||

3 Implications

We now establish some characteristics of equilibrium behavior. When considering the comparative statics of $\lambda = x_2/(x_1+x_2)$, we keep the total x_1+x_2 constant throughout. We rely heavily on the presumptions that utilities are homogeneous of degree one in monetary payoffs and that the perturbations

Θ_C and Θ_D are small.

Let $\hat{\Delta}_1(\alpha, \lambda)$ and $\hat{\Delta}_2(\alpha, \lambda)$ be the equilibrium functions $\Delta_1(\alpha)$ and $\Delta_2(\alpha)$ (cf. Proposition 2) in the maximally cooperative equilibrium given λ. Let an outcome of the two-period game be written as (for example) (CD, DD), in which case player 1 cooperated and 2 defected in the first period, with both players defecting in the second period.

Proposition 3 *Let Assumptions 1–4 hold and let $\lambda \in (0, 1)$ unless otherwise stated. Then, for sufficiently small Θ_C and Θ_D:*

(3.1) In any monotonic equilibrium with $\lambda \in (0, 1)$, the expected incidence with which a player cooperates, where the expectation is taken over the values of α, Θ_C and Θ_D and over the opponent's behavior, is higher in the first period than in the second.

(3.2) Equilibrium play in the first period of a game with $\lambda = 0$ is identical to equilibrium play in the second period of a game with $\lambda = 1$.

(3.3) Maximally cooperative monotonic equilibria satisfy

$$\lambda' > \lambda \quad \Rightarrow \quad \hat{\Delta}_1(\alpha, \lambda') < \hat{\Delta}_1(\alpha, \lambda)$$

for $\lambda, \lambda' \in [0, 1)$. Hence, the larger is λ, the more likely is a player to cooperate in the first period.

(3.4) As the value of $\lambda \in (0, 1)$ increases, the expected incidence of outcomes (CD, CD), (DC, DC), and (DD, DD) in a maximally cooperative monotonic equilibrium decline. The expected incidences of (CC, DD), (CC, CD) and (CC, DC) are approximately zero for small values of λ (when $\Theta_C = \Theta_D = 0$, values of λ for which $\alpha_1 < 1$), with (CC, DD) thereafter increasing in λ and (CC, CD) and $CC, DC)$ thereafter positive but with an ambiguous comparative static in λ. Expected monetary payoffs from a maximally cooperative monotonic equilibrium first increase in λ, reach an interior maximum, and then decrease.

The first result indicates that we expect the incidence of cooperation to decrease as we reach the end of the finite-horizon of play. The incentive to cooperate in the first period varies as does λ, but is greater than the incentive to cooperate in the second period for every value $\lambda \in (0, 1)$.[10] Hence, the specter of the future enhances current cooperation. Notice that this result is not simply part of the definition of a monotonic equilibrium. Monotonicity indicates that the incidence of second-period cooperation is increasing in first period cooperation, but says nothing about the relative magnitudes of first-period and second-period cooperation.

[10] When $\lambda = 0$ or $\lambda = 1$, one period is irrelevant, in which play is arbitrary.

The second result indicates that if all monetary payoffs are concentrated in a single period, then we can expect identical play (and hence expected payoffs) whether that period is the first or second. When $\lambda = 0$, the irrelevance of second-period play ensures that play in the first-period must match that of the unique equilibrium of the one-stage game. When $\lambda = 1$, first-period actions are irrelevant and the equilibrium of the one-stage game appears in the second period.

Statement (3.3) indicates that first-period cooperation increases as the second period becomes more important. As the stakes are shifted to the second period, players are increasingly willing to invest their first-period behavior in encouraging second-period cooperation.

The final statement provides predictions for those two-period outcomes for which results are unambiguous and then derives the effects of these behavioral shifts on payoffs.[11] As λ increases, first-period cooperation increases, while payoffs are shifted to the second period, where some defection occurs. The result is that overall payoffs are first increasing and then decreasing in λ, finding their maximum when second-period payoffs are sufficiently important, but not arbitrarily larger than first-period payoffs. Though we cannot make concrete statements without additional information about the specification of preferences and the distribution F, our expectation is that expected payoffs will be maximized when $\lambda > \frac{1}{2}$, in which case the relationship begins with relatively small stakes and works up to larger stakes. This will be the case, for example, if F is uniform.

Proof. We first describe monotonic equilibria in the limiting case in which $\Theta_C = \Theta_D = 0$. To do so, let

$$H : [\underline{\alpha}, \overline{\alpha}]^2 \rightarrow \{(z_1, z_1', z_2, z_2') | z_1, z_1', z_2, z_2' \in \{C, D\}\}$$

be a function with the property that $H(\alpha, \alpha')$ identifies the equilibrium path of play given that the players' actual types are α and α'. Hence, $H(\alpha, \alpha') = \{DC, DD\}$ indicates that players α and α' defect and cooperate (respectively) in the first period and that both defect in the second period. Let $p_i(H(\alpha, \alpha'))$ be the equilibrium period-i monetary payoff of agent α when paired with agent α'. Then expected equilibrium payoffs are given by

$$\int_{\alpha'} \int_{\alpha} (p_1(H(\alpha, \alpha')) + p_2(H(\alpha, \alpha'))) dF(\alpha) dF(\alpha').$$

[11] We establish one more unambiguous finding, that the expected incidence of (CC, CC) increases if $\alpha_1(\lambda) < 1$ and decreases if $\alpha_1(\lambda) > 1$, but the outcome (CC, CC) appears too seldom in our data to evaluate this result.

Figure 6 illustrates the function $H(\alpha, \alpha')$ for an equilibrium in which $\alpha_1 < 1$ and an equilibrium in which $\alpha_1 > 1$.

(3.1) Figure 6 shows that when $\Theta_C = \Theta_D = 0$, then for each possible outcome path in a monontonic equilibrium, there is weakly more cooperation in the first period than the second, with a strict inequality for some values of (α, α'). Taking an expectation over (α, α'), we thus get a higher expected incidence of cooperation in the first period. The upper hemicontinuity of the equilibrium correspondence at $\Theta_C = \Theta_D = 0$ extends this to small values of $\Theta_C = \Theta_D$.

(3.2) The uniqueness of the one-shot equilibrium and the irrelevance of second-period behavior when $\lambda = 0$ implies that $\hat{\Delta}_1(\alpha, 0)$ is unique and equals $\Delta^*(\alpha)$. Next, letting $\lambda = 1$, we can construct an equilibrium of the two-period game in which players choose identical (possibly mixed) actions in the first stage, which are then uninformative, with the unique one-shot equilibrium appearing in the second stage after every first-period outcome. The fact that all players prefer increased cooperation on the part of their opponents ensures that there are no other equilibria. In particular, if first-period plays of C and D gave rise to different probabilities of second-period cooperation, all agents would choose the first-period action giving the highest probability of second-period opponent cooperation, rendering first-period actions uninformative.

(3.3) To establish the third result, fix λ and let Θ_C and Θ_D be sufficiently small that the second-period equilibrium is unique. Consider player $\alpha_1(0, \lambda)$, defined to be the player who is indifferent between C and D in the first period of a maximally cooperative equilibrium, given λ and given $\delta = 0$. Notice that the value $\alpha_1(0, \lambda)$ uniquely determines the function $\hat{\Delta}_1(\alpha, \lambda)$, in the sense that knowing the value of $\alpha_1(0, \lambda)$ gives us enough information to calculate the remainder of the function. Lemma 2 implies that

$$\begin{aligned} \pi(C, \alpha_1(0,\lambda), \alpha_1(0,\lambda)) &< \pi(D, \alpha_1(0,\lambda), \alpha_1(0,\lambda)) \\ V(C, \alpha_1(0,\lambda)) &> V(D, \alpha_1(0,\lambda)). \end{aligned}$$

Using the assumption that π is linearly homogeneous in monetary payoffs, an increase from λ to λ' then causes this player to strictly prefer C in the first period. Hence, there exists a set $[\alpha_1(0, \lambda), \hat{\alpha}] \subset [\alpha_1(0, \lambda), \overline{\alpha}]$ with the property that given the second-period strategies prescribed by Proposition 2 for any given value of $\alpha_1 \in [\alpha_1(\lambda, 0), \hat{\alpha}]$ and given λ', player α_1 at least weakly prefers C in the first period. Now there are two possibilities. First, if $\hat{\alpha} < \overline{\alpha}$, then $\hat{\alpha}$ must be indifferent between C and D in the first period (otherwise $\hat{\alpha}$ would not be an upper bound), and then we would have an

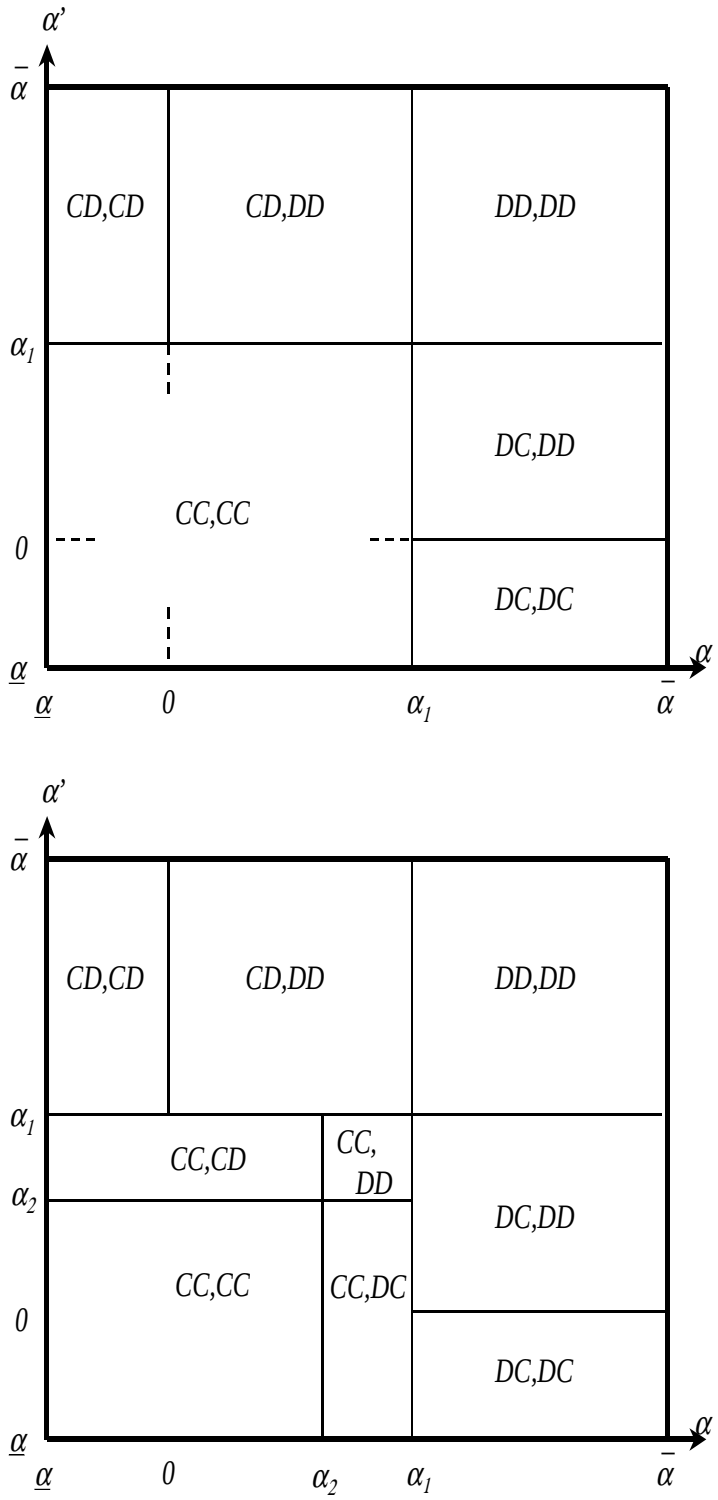

Figure 6: *Function $H(\alpha, \alpha')$ corresponding to a monotonic equilibrium, where $\alpha_1 < 1$ (top) and $\alpha_1 > 1$ (bottom).*

equilibrium given λ' (obtained by letting $\alpha_1(\lambda', 0) = \hat{\alpha} > \alpha_1(\lambda, 0)$) featuring more cooperation than the maximally cooperative equilibrium given λ, establishing the result. Second, if $\hat{\alpha} = \overline{\alpha}$, then there is an equilibrium with $\alpha_1(\lambda', 0) = \overline{\alpha} > \alpha_1(\lambda, 0)$, again establishing the result.

(3.4) We first establish the final result for the case of $\Theta_C = \Theta_D = 0$. Using the top panel in Figure 6 as a guide, consider an increase in λ and hence α_1 (from Proposition 3.3), beginning with the minimum value of $\alpha_1(0) = \alpha^* < 1$ (the latter because $F(1) < 1$). As α_1 increases, we have the following possible transitions in behavior, each applicable to some values of (α, α'):

$$\begin{array}{rcl} (CD, CD) & \rightarrow & (CC, CC) \\ (CD, DD) & \rightarrow & (CC, CC) \\ (DC, DC) & \rightarrow & (CC, CC) \\ (DC, DD) & \rightarrow & (CC, CC) \\ (DD, DD) & \rightarrow & (CC, CC) \\ (DD, DD) & \rightarrow & (CD, DD) \\ (DD, DD) & \rightarrow & (DC, DD). \end{array} \tag{17}$$

Hence, the incidence of (CD, CD), (DC, DC) and (DD, DD) falls, while (CC, CC) increases.

Next, suppose that α_1 reaches 1, so that the bottom panel of Figure 6 is relevant. It follows from (15) that $\alpha_2 = \alpha_1$ when $\alpha_1 = 1$, with α_2 decreasing as α_1 increases above 1. We then have the following list of possible transitions:

$$\begin{array}{rcl@{\qquad}rcl} (CD, CD) & \rightarrow & (CC, CD) & (DD, DD) & \rightarrow & (CD, DD) \\ (CD, DD) & \rightarrow & (CC, CD) & (DD, DD) & \rightarrow & (DC, DD) \\ (CD, DD) & \rightarrow & (CC, DD) & (CC, CC) & \rightarrow & (CC, DD) \\ (DC, DC) & \rightarrow & (CC, DC) & (CC, CC) & \rightarrow & (CC, CD) \\ (DC, DD) & \rightarrow & (CC, DC) & (CC, CC) & \rightarrow & (CC, DC) \\ (DC, DD) & \rightarrow & (CC, DD) & (CC, CD) & \rightarrow & (CC, DD) \\ (DD, DD) & \rightarrow & (CC, DD) & (CC, DC) & \rightarrow & (CC, DD). \end{array} \tag{18}$$

Again, the conclusion is that the incidence of (CD, CD), (DC, DC) and (DD, DD) falls. Hence, each of these three outcomes declines as λ increases. However, we now also find that (CC, CC) decreases, leading to the observation that (CC, CC) increases for λ such that $\alpha_1(0, \lambda) < 1$ and decreases for λ such that $\alpha_1(0, \lambda) > 1$. Next, (18) shows that (CC, DD) increases in λ for $\alpha_1(0, \lambda) > 1$, while (CC, DD) does not appear when $\alpha_1(0, \lambda) < 1$.

Finally, note that (CC, DC) and (CC, CD) appear only for values of λ with $\alpha_1(0, \lambda) > 1$, and, from (18), that there are transitions both to and from these outcomes as λ increases when $\alpha_1(0, \lambda) > 1$, leading to an ambiguous comparative static.

Now return to the transitions in (17). In each case, the monetary payoff realized by the pair (α, α') increases in each period, and hence so must the total monetary payoff increase. Hence, the equilibrium expected monetary payoff, with the expectation taken over the values of (α, α'), increases in λ, and will be increasing in λ until at least $\alpha_1(0, \lambda) = 1$. At this point, the transitions in (18) become relevant. Every transition again enhances the incidence of first-period cooperation, so that further increases in λ and hence $\alpha_1(0, \lambda)$ will increase first-period cooperation. However, a conflicting force arises, with some transitions now *decreasing* second-period cooperation (e.g., $(CC, CC) \rightarrow (CC, DD)$, $(CC, CC) \rightarrow (CC, CD)$ and $(CC, CC) \rightarrow (CC, DC)$).[12] Monetary payoffs are thus shifted away from the enhanced first-period cooperation to a second period in which the incidence of defection is increasing. Eventually, this latter force will dominate, causing the expected payoff to fall. In particular, the expected payoff corresponding to $\lambda = 1$ equals that corresponding to $\lambda = 0$, with higher payoffs for intermediate values.

The argument extends to cases in which Θ_C and Θ_D are nonzero but small, upon noting that behavior in such equilibria approaches that shown in Figure 6 as Θ_C and Θ_D approach zero. ||

4 Experimental Procedures

The experiment was conducted at the University of Wisconsin, using undergraduate subjects, in May and October of 2002, in five sessions involving 22 subjects each. Each session involved 20 "rounds," in each of which all subjects in that session were matched with opponents from their session for a twice-played or "two-period" prisoners' dilemma. Hence, in each of the 20 rounds the 110 subjects were matched in 55 pairs to play the two-period game, for a total of 1100 two-period games.

Subjects interacted via an anonymous computer interface. Subjects were randomly matched with a partner for each round subject to the constraint that no subject played the same opponent more than once. These details

[12] This occurs because a first-period observation of (C, C) is now followed by second-period strategies of cooperating if and only if $\alpha < \alpha_2 < \alpha_1$ (cf. (13)).

λ	Period-1		Period-2		Frequency
	Pull value	Push value	Pull value	Push value	
0.0	10	30	0	0	101
0.1	9	27	1	3	110
0.2	8	24	2	6	102
0.3	7	21	3	9	85
0.4	6	18	4	12	90
0.5	5	15	5	15	91
0.6	4	12	6	18	115
0.7	3	9	7	21	112
0.8	2	6	8	24	95
0.9	1	3	9	27	94
1.0	0	0	10	30	105
					1100

Figure 7: *Values of λ, with the corresponding period-1 pull value x_1 (cf. Figure 1), period-1 push value $3x_1$, period-2 pull value $x_2 = 10 - x_1$, and period-2 push value $3x_2$, and the number of two-period games (out of 1100) corresponding to each value.*

were known to the subjects.[13]

The prisoners' dilemma was presented to the subjects as the *push-pull* game. In each stage game, each subject had the opportunity to either *pull* x points toward the subject or *push* $3x$ points toward the opponent. Each subject earned the total of whatever sum they pulled and their opponent pushed, leading to the payoffs of Figure 1. We regard this as a particularly simple way of presenting the prisoners' dilemma.

Let the period-1 and period-2 pull values be denoted by x_1 and x_2. In each two-period game, $x_1 + x_2 = 10$. The value of $\lambda = \frac{x_2}{x_1+x_2}$ was randomly selected every time a pair of subjects was matched to play a two-period game, independently across pairs of subjects and games, with λ drawn from the set $\{0, 0.1, 0.2, \ldots, 1\}$. Table 7 shows the distribution of realized values of λ in the experiment.

Subjects were paid their cumulative earnings, in cash, at the end of the experiment. The pull and push values x_i and $3x_i$ identified points that the subject earned in each game. In two of the five sessions, each point was

[13] Instructions were provided to the subjects via computer. The instructions are available at http://www.ssc.wisc.edu/ larrysam/extras.htm.

worth two cents and subjects were also paid a five-dollar show-up fee, with earnings ranging from $8.00 to $12.48 for an experiment that lasted less than an hour. In three of the five sessions, each point was worth six cents (with no show-up fee), with earnings ranging from $7.02 to $24.00. We found no significant differences in behavior between the two payment schemes.[14]

5 Results

This section provides a summary of the experimental outcomes and then examines each of the four parts of Proposition 3. Our working hypothesis, which we refer to as the *equilibrium hypothesis*, is that the play of the experimental subjects is described by an equilibrium of the game. In particular, we assume that each player is characterized by a realized value of α that remains fixed throughout the experimental session, where these values are independently drawn from a distribution F. We assume that in each period of each two-period game, each player (independently) draws realizations θ_C and θ_D of the random variables Θ_C and Θ_D. Given these realizations, we assume that the subjects play their part of the equilibrium described in Propositions 1–3.

5.1 Summary of Outcomes

Each subject had 40 opportunities to either push (cooperate) or pull (defect), one in each of the two periods of twenty games. Figure 8 reports the distribution across subjects of the overall incidence of cooperation.

Figure 8 shows that only two subjects cooperated more than 30 out of 40 times (cooperating 31 and 40 times). Since a player for whom $\alpha < 0$ is predicted to cooperate at every opportunity (in the limit as the noise level gets small), our results are thus consistent with the players having been drawn from a distribution F for which $F(0)$ is small. This in turn suggests that equation (10) of Assumption 4 is reasonable. Whereas there is no subject who defects at every opportunity, there are more subjects who come closer to persistent defection than to persistent cooperation. Hence, equation (11) of Assumption 4 may also be reasonable.

Figure 9 identifies the outcomes of the 1100 two-period games. Given that each player has two choices in each of two periods, there are sixteen possible paths of play. However, we are not interested in distinguishing

[14]Because no subject participated in more than one session, the individual fixed effects in our regressions below provide a control for session differences.

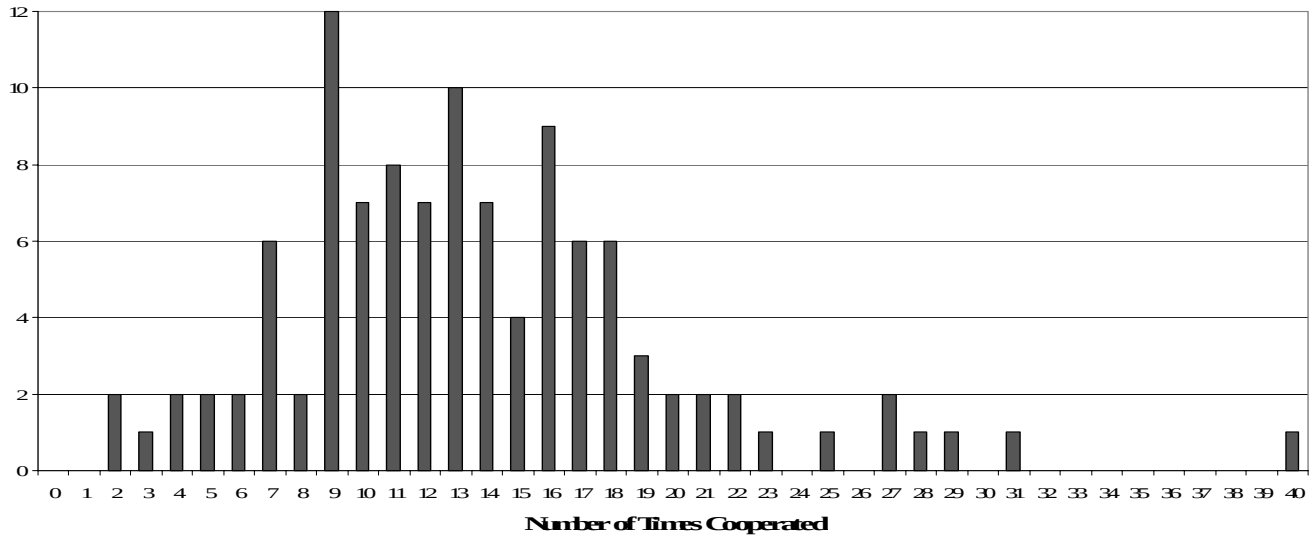

Figure 8: *Each subject faced 40 cooperate/defect decisions. The histogram identifies the number of subjects (vertical axis) exhibiting each of these possible frequencies of cooperation (0 to 40, on the horizontal axis).*

outcomes that are identical except for which player is labeled "player 1" and which "player 2," allowing us to collapse these outcomes to ten cases. Hence, the last line of Figure 9 corresponds to a case in which one player defected in period 1 (whom we have designated player 1) and one cooperated, while both cooperated in period 2.[15] Figure 9 presents data for all games as well as for those games in which $\lambda \in (0, 1)$, so that payoffs are relevant in both periods. These are the games relevant to many of the comparative static predictions in Proposition 3.[16]

Figure 9 indicates that an agent's opponent is more likely to cooperate in the second period when the agent cooperates in the first period. Restricting

[15] To construct Figure 9, we chose one player in each pair to be player 1 and then list player 1's action first in both periods. If only one player defected in period one, that player was chosen to be player 1. If the players chose the same action in period 1 and only one defected in period 2, that player was chosen to be player 1. The outcomes that could be coded as (DC, DC) and (CD, CD) are thus effectively identical, differing only in the identify of the player chosen to be named player 1, and both are represented as (DC, DC). The outcomes (DC, CD) and (DC, DC) are different, since the two players simply repeated the first-period actions in the second period of the second case, but switch actions in the second period of the first case.

[16] Lines 5, 6, 9, and 10 of Figure 9 correspond to outcomes not predicted by a noiseless version of the model.

Period one	Period two	Frequency	Frequency, $\lambda \in (0,1)$
CC	DD	191	127
CC	DC	90	79
CC	CC	30	26
DD	DD	242	234
DD	DC	121	90
DD	CC	60	19
DC	DD	263	239
DC	DC	42	36
DC	CD	43	33
DC	CC	18	11
		1100	894

Figure 9: *Frequency of each possible outcome for the two-period game. "DC CD," for example, indicates that one player defected and one cooperated in the first period, and then the two players switched actions for the second period.*

attention to cases in which $\lambda \in (0,1)$, we have:

Cooperative opponent play in period 2,
given agent cooperation in period 1 : 22% (175 of 783)

Cooperative opponent play in period 2,
given agent defection in period 1 : 17% (175 of 1005).

These data are consistent with the presumption that the equilibrium is monotonic.

5.2 Proposition 3.1: Cooperation by Period

Figure 9 allows us to assess Proposition 3.1, asserting that cooperation will be more prevalent in the first period when $\lambda \in (0,1)$. We have:

Cooperative plays in period 1 : 44% (783 of 1788)
Cooperative plays in period 2 : 20% (350 of 1788).

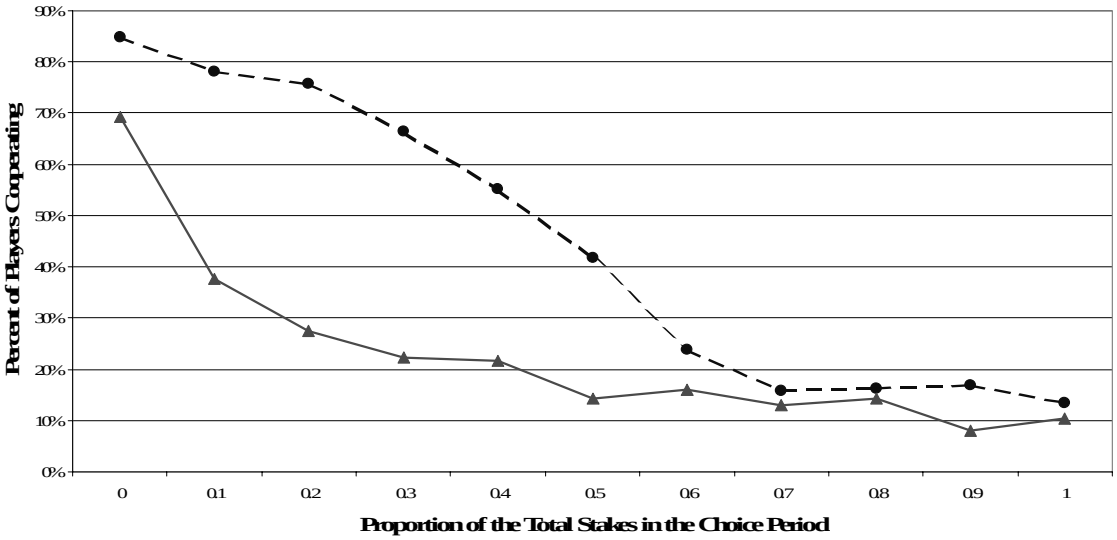

Figure 10: *Incidence of first-period (dashed line) cooperation and second-period (solid line) cooperation, each as a function of the proportion of the total stakes in that period. This groups together first-period choices and second-period choices made with the same stakes in the choice period. Proposition 3.1 predicts there will be more first-period cooperation when* $\lambda \in (0, 1)$.

Figure 10 illustrates this result, showing the incidence of first-period cooperation and the incidence of second-period cooperation, each as a function of the proportion of the total stakes in that period. Thus at .4, we compare the first-period play for $\lambda = .6$ (forty percent of the stakes in the first period) with second-period play for $\lambda = .4$ (forty percent of the stakes in the second period).

Figure 10 shows the predicted result that cooperation is more prevalent in the first than in the second period, especially when the first-period stakes are small.[17] It is intuitive that the difference between first and second-period cooperation would be greatest when first-period stakes are small, since it is

[17] To obtain a more formal assessment of the significance of this difference, let η_{ij} be the number of times subject i cooperated in period j. Under the equilibrium hypothesis, we can view each η_{i1} as the result of 20 independent Bernoulli trials whose mean is unknown and idiosyncratic to player i (depending upon i's draw of the parameter α). However, η_{i2} is not independent of η_{i1}. Let $\eta_i = \eta_{i1} - \eta_{i2}$ be the difference in the number of times subject i cooperated in the first period and the second period. Let $\overline{\eta}$ and s_η be the mean and standard deviation of the sample values η_i, $i = 1, \ldots, 110$. Given our sample size of 110, we can assume that the distribution of $\overline{\eta}/(s_\eta/\sqrt{110})$ is approximately Normal with zero mean and unitary variance (given the hypothesis that there is no difference in the incidence of first-period and second-period cooperation). We find:

	Period one, $\lambda = 0$	Period two, $\lambda = 1$
Number of cooperators:	27	22
Percentage:	13.4%	10.5%
Number of defectors:	175	188
Percentage:	86.6%	89.5%

Figure 11: *Comparison of play in first period of games with irrelevant second-period payoffs ($\lambda = 0$) and second period of games with irrelevant first-period payoffs ($\lambda = 1$). Proposition 3.2 predicts the same behavior in these two games.*

here that the cost of investing in second-period cooperation is especially low.

5.3 Proposition 3.2: Effectively Single-Shot Games

Proposition 3.2 indicates that we should expect play in the first period of games with $\lambda = 0$ to be identical to play in the second period of games with $\lambda = 1$. In each case, the two-period game includes one stage game played for zero payoffs, with all of the payoffs concentrated in the other game. They differ in whether the zero-payoff period comes first ($\lambda = 1$) or second ($\lambda = 0$).

Figure 11 presents play in the first period of games with $\lambda = 0$ and in the second period of games with $\lambda = 1$. Play in these two games is similar. To obtain a more precise estimate of this similarity, let η_{i1} be the percentage of the time that player i cooperated when playing in period 1 of a game with $\lambda = 0$ and let η_{i2} be the percentage of the time that player i cooperated when playing in period 2 of a game with $\lambda = 1$. Under our hypotheses and experimental design, these percentages represent the outcomes of draws from independent Bernoulli trials (with different numbers of draws for different subjects, since the value of λ is drawn randomly in each game). Our model predicts that the two random variables pertaining to any particular subject will have the same mean, though these means can differ across subjects. Let

$\overline{\eta}$	s_η	$\overline{\eta}/(s_\eta/\sqrt{110})$	p-value
4.33	4.18	10.9	0.000

,

where the p-value gives the (approximate) probability of generating a test statistic whose value exceeds 10.9 under the null hypothesis of no difference in play across periods.

$\eta_i = \eta_{i1} - \eta_{i2}$, for those 80 subjects who faced games with both $\lambda = 0$ and $\lambda = 1$. Letting $\overline{\eta}$ and s_η be the mean and standard deviation of the observed values of η_i, the distribution of $\overline{\eta}/(s_\eta/\sqrt{80})$ is approximately Normal with zero mean and unitary variance. Then we calculate:

$\overline{\eta}$	s_η	$\overline{\eta}/(s_\eta/\sqrt{80})$	p-value
2.3	28.3	0.71	0.48

where the p-value is the probability of obtaining a coefficient on the variable "period two" whose absolute value exceeds 0.71, given the null hypothesis of no difference between periods.[18]

The second period thus has an effect on cooperation that is statistically small. Section 5.5 shows that the induced difference in payoffs is also small. Though the behavior does not match exactly, the subjects in our experiments appear to recognize circumstances under which only one period of the game is relevant and to treat that period similarly, whether it occurs as the first or second stage of the two-period prisoners' dilemma.

5.4 Proposition 3.3: First-Period Cooperation

Proposition 3.3 predicts that the incidence of first-period cooperation should be increasing in λ when $\lambda \in [0, 1)$. The higher is λ, the more important are second-period payoffs, and hence the more valuable it is to invest in second-period cooperation by cooperating in the first period. Figure 12 shows first-period cooperation as a function of λ. The incidence of cooperation increases from 13% when $\lambda = 0$ to 78% when $\lambda = .9$.

To confirm this link between the concentration of payoffs in the second period and first-period cooperation, we examine a logit regression in which the dependent variable equals one if a player cooperated in period 1 and 0 otherwise. There are 1990 observations, one for each choice by one of the 110 agents in the first period of each game played by the agent in which $\lambda \in [0, 1)$. Independent variables include λ, a constant, and nineteen dum-

[18] This calculation ignores data from the 16 subjects who participated only in games with $\lambda = 1$ (who cooperated 11.4% of the time in the second period such games) and from the 13 subjects who participated only in games with $\lambda = 0$ (who cooperated 16.2% of the time in the first period of such games). (One subject participated in no games with $\lambda = 0$ or $\lambda = 1$.) These data can be included, at the cost of making an assumption concerning homogeneneous behavior across players that we prefer to avoid, yielding an estimate of the difference between the first-period $\lambda = 0$ case and the second-period $\lambda = 1$ case that is again not significant.

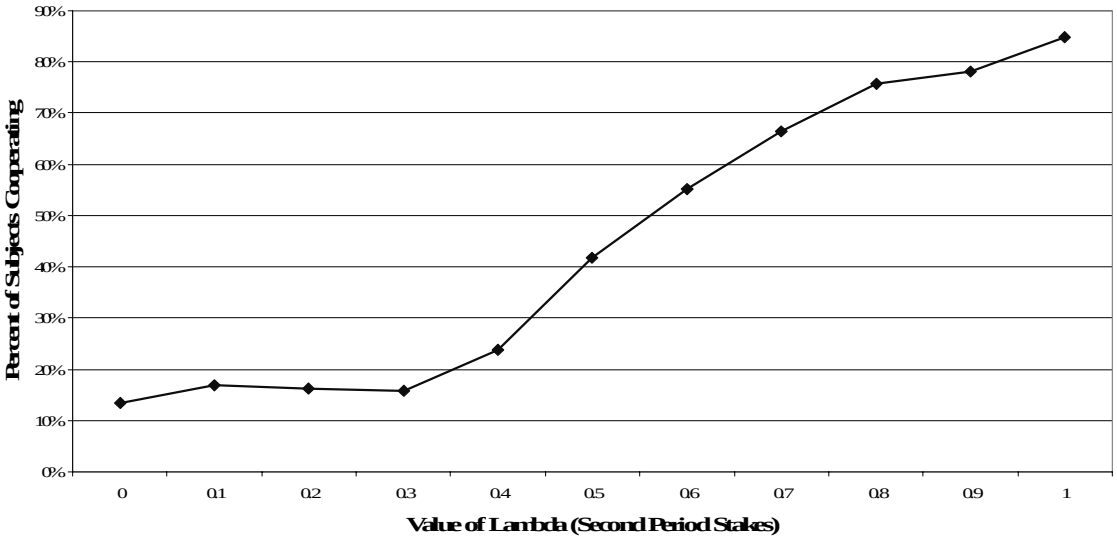

Figure 12: *First-period cooperation as a function of the importance of second-period payoffs, ranging from games in which second-period payoffs are irrelevant ($\lambda = 0$) to games in which only second-period payoffs matter ($\lambda = 1$). Proposition 3.3 predicts that the incidence of cooperation will increase in λ for $\lambda \in [0, 1)$.*

mies identifying the round (from 1 to 20) of the observation.[19] Our model, in conjunction with the equilibrium hypothesis, indicates that first-period behavior in each game can be viewed as a draw from an independent random variable with a mean that is potentially idiosyncratic to the subject in question. To incorporate the correlations introduced by the dependence between multiple observations on the part of a single player, we include fixed effects for the players in the regression.[20] Our interest centers on the estimated

[19]This allows the possibility that increased subject familiarity with the game may cause behavior to differ across rounds, though our equilibruim hypothesis is that such considerations are of secondary importance. Subjects tend to be somewhat less cooperative in later rounds, but respond to λ consistently throughout the experiment.

[20]Under our equilibrium hypothesis, the fixed effects appropriately incorporate the correlations between the multiple observations of play on the part of a single player. If the equilibrium hypothesis fails, a player's actions in round t of the experiment may depend upon her experience in rounds $1, \ldots, t-1$. It would still be appropriate to omit this history from the regression, and the fixed effects would adequately capture the dependence between different observations from a single player, as long as the history observed by a player is not correlated with the fixed-effects player-specific error term. Such a correlation could appear, as player i's play could affect the subsequent behavior of the current opponent j, who might meet and effect the subsequent behavior of a player k who subsequently

coefficient for λ. Suppressing the other estimates, we find:

Variable	Estimated coefficient	Standard error	p-value
λ	6.1	.31	0.000

The p-value is the probability of generating a coefficient estimate whose value exceeds 6.1 given that the null hypothesis of a zero coefficient. As expected, λ is significantly positive: higher period-two payoffs yield higher first-period cooperation.

5.5 Proposition 3.4: Expected Payoffs

Proposition 3.4 begins with predictions concerning equilibrium outcomes. First, the outcomes (CC, DD) and (CC, DC) are predicted to be rare (i.e., to not occur when the random payoff perturbations θ_C and θ_D are zero) for relatively small values of λ, with (CC, DD) increasing for larger values of λ and with (CC, DC) positive but with an ambiguous comparative static for these larger values.[21] Figure 13 plots the incidence of these outcomes as a function of λ. The predicted trends are clearly visible. The common value of λ above which these outcomes appear more frequently, corresponding to $\alpha_1 = 1$, is predicted to be below the value of λ that maximizes the sum of the two-players' expected payoffs, which is consistent with the findings reported below.

The incidences of (DC, DC) and (DD, DD) are predicted to decrease in λ for $\lambda \in (0, 1)$. Figure 14 shows the relevant results. Small sample sizes obscure a hint of a downward trend in the incidence of (DC, DC) as λ rises (note the difference in vertical scales in Figures 13 and 14). Once we eliminate the value corresponding to the incidence of (DD, DD) when $\lambda = 0$, *for which the model makes no prediction*, it is clear that (DD, DD) appears less often for higher values of λ, though the realizations are not perfectly monotonic. Each of the three "naked eye" relationships for which sample sizes are adequate can be confirmed by regressions, but we omit the details.

Proposition 3.4 next suggests that total payoffs from the two-period game should be first increasing in λ and then decreasing in λ, attaining an interior maximum. The cases of $\lambda = 0$ and $\lambda = 1$ should give equal payoffs,

encounters i. (Recall that no two players ever meet more than once.)

[21] Note that the outcomes (CC, CD) and (CC, DC) that appear in Proposition 3.4 are observationally equivalent (cf. Figure 9), as are the outcomes (CD, CD) and (DC, DC), discussed below.

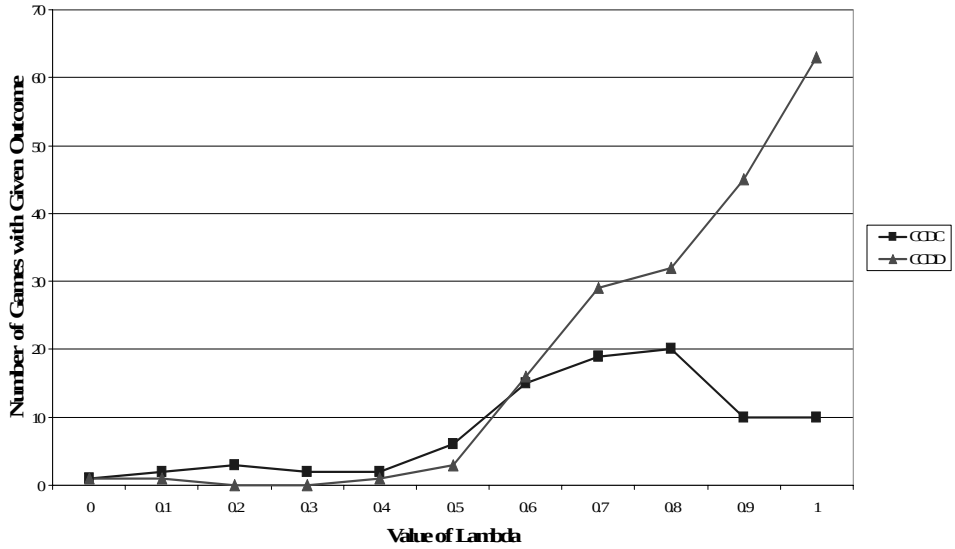

Figure 13: *Frequency of outcomes* (CC, DD)*, and* (CC, DC) *as a function of* λ*. Proposition 3.4 predicts the incidence of* (CC, DD) *and* (CC, DC) *will be approximately zero over a common range of relatively small values of* λ*, above which* (CC, DD) *is increasing and* (CC, DC) *is positive but with an ambiguous comparative static.*

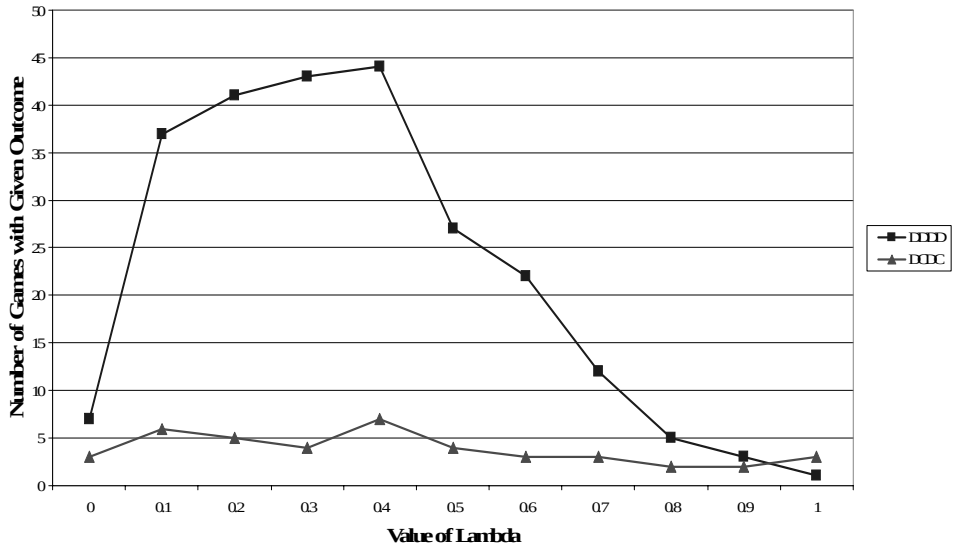

Figure 14: *Frequency of outcomes* (DD, DD) *and* (DC, DC) *as a function of* λ*. Proposition 3.4 makes no prediction for* $\lambda = 0$*, and predicts that both will fall as* λ *increases for those values of* $\lambda \in (0, 1)$*.*

Value of λ	Average payoff
0.0	25.4
0.1	27.5
0.2	27.4
0.3	27.1
0.4	29.2
0.5	31.2
0.6	32.7
0.7	31.6
0.8	30.6
0.9	26.0
1.0	24.2

Figure 15: *Average value of total payoffs for each value of λ. "Total payoff" is the sum of the period-one and period-two payoffs earned by both players in a two-period game, measured in points. The number of observations for each value of λ is given in Figure 7.*

reflecting the equivalent behavior for $\lambda = 0$ and $\lambda = 1$. Figure 15 shows the average earnings, summed over players and periods, for each value of λ. The minimum possible payoff is 20 (experimental points) and the maximum is 60, with every even number between these two values being possible and with every such value occurring in the data. The maximum observed payoff occurs at $\lambda = 0.6$, with a payoff 35% higher than the minimum payoff. The payoffs attached to $\lambda = 0$ and $\lambda = 1$ are quite similar.

To assess the relationship between λ and total payoffs , we turn to a regression in which the dependent variable is the total (over players and periods) of the number of experimental points earned in the two-period prisoner' dilemma. Independent variables include λ, λ^2, and λ^3 as well as a constant, individual fixed effects and dummy variables identifying the round of the experiment in which the game is played.[22] There were 1100 observations, one for each of the 11 two-period games played in each of 20

[22]Our model tells us that we require at least a cubic equation to examine this relationship. The model predicts that the relationship should be nonmonotonic, that the values $\lambda = 0$ and $\lambda = 1$ give equal total payoffs, and that total payoffs should achieve an interior maximum that we suspect will occur at a value $\lambda > \frac{1}{2}$. A quadratic equation would suffice to allow nonmonotonicity, but can equate the total payoffs at $\lambda = 0$ and $\lambda = 1$ only if it forces the maximum to occur at $\lambda = \frac{1}{2}$. Investigating the value of the λ that maximizes total payoffs thus requires a cubic equation.

rounds in each of 5 experimental sessions. We are interested in the estimated coefficients on λ, λ^2 and λ^3, for which we find:

Variable	Estimated coefficient	Standard error	p-value
λ	-4.5	6.7	0.50
λ^2	51	16	0.001
λ^3	-49	10	0.000

The coefficient on λ is negative, relatively small, and insignificant. The coefficients on λ^2 and λ^3 are much larger in absolute value, of opposite signs, and both significant. They combine for an estimated relationship that is initially increasing, reaches a maximum at $\lambda = .65$, and then decreases.

Expected payoffs are thus nonmonotonic in λ. If one's goal is to maximize the expected total payoff generated by a two-period prisoners' dilemma, then one should neither pack all of the payoffs into the first period nor into the second period. Instead, doing so would *minimize* the expected payoff, with very little depending upon which period contained the relevant payoffs and which was irrelevant. Payoffs are maximized by making the stakes in period two between one-and-one-half and two times as large as those in period one.

6 Discussion

Summary. The point of departure for our research is the experimental evidence that individuals have tastes for cooperation in the prisoners' dilemma that, at times, override economic incentives to the contrary. We view our research as a first step toward understanding how formal or informal institutions might be designed to utilize these private tastes in order to facilitate more efficient economic and social interactions.

The keys to our model of behavior in the prisoners' dilemma are the hypotheses that people prefer that their opponents cooperate, sometimes prefer to cooperate themselves, may differ in the strength of this preference, and value cooperation relatively more when their opponents are likely to cooperate. The first three hypotheses appear to be essential elements of any model of prisoners' dilemma behavior. Our analysis becomes nontrivial when adding the final hypothesis and identifying implications for how behavior should be affected by changing the relative magnitudes of payoffs across the periods of a twice-played prisoners' dilemma. Here, we obtain some predictions that are not immediately obvious, including the increase in first-period cooperation as second period payoffs become relatively important and especially the interior maximum of expected total payoffs as a function of

the relative importance of second-period payoffs. These predictions appear clearly in the data, leading us to believe that our model captures some robust features of behavior.

Extensions. There are many more avenues to explore when asking how to construct social and economic interactions to take advantage of natural inclinations to cooperate, whether arising out of feelings of altruism, fairness or trust. In particular, we think it important to extend our analysis beyond the two-period prisoners' dilemma game considered here.[23] For example, we would be interested in whether starting small has more of an effect in longer games. The length of the game as well as the distribution of stakes could become a feature of the institutional design.

We are also interested in applying similar ideas to the "trust game." Player 1 in the trust game is endowed with a sum of money that she can divide between herself and player 2. Any money given to player 2 is tripled, with player 2 then dividing the resulting sum between player 1 and himself. The subgame perfect equilibrium calls for player 1 to donate none of the money to player 2, while the efficient outcome calls for player 1 to donate all of the money to player 2. Experimental results show that player 1 typically donates some (but not all) of the money to player 2, who on average offers a return that does not fully compensate player 1 for the donation.

Again, we suspect that trust can be nurtured in repeated play of this game by appropriately choosing the size of the stakes. Examining this and similar games will contribute to a better understanding of the interaction between interperiod payoff variation and cooperation or trust. Combined, these studies will help develop an understanding of how simple institutions can be designed to foster and exploit natural tastes for cooperation.

7 Appendix

Proof of Proposition 2.1: Uniqueness. To establish uniqueness, we must consider continuation play when one agent has cooperated and one has defected in period one. Assumption 1 ensures that there exist α' and α'', with

$$0 \leq \alpha' \leq \alpha_1 \leq \alpha'' \leq 1,$$

with the player α who cooperated in period one (in the probability-one event that $\alpha \in [\underline{\alpha}, \alpha_1]$) cooperating in the second period if and only if $\alpha < \alpha'$; and

[23] This paper focuses on two-period games because comparative static results are much more difficult to extract from longer games.

the first-period defector α (in the probability-one event that $\alpha \in [\alpha_1, \overline{\alpha}]$) cooperating in period two if and only if $\alpha < \alpha''$. This is simply the statement that, conditional on first-period behavior, players with relatively low values of α will cooperate and those with high values of α will defect (coupled with the observation that agents with dominant second-period strategies will play them).

Now consider the possibilities for the value of α'. The penultimate paragraph of the proof of Proposition 2.1 in the text showed, with the help of condition (10), that $\alpha' = 0$ implies that $\alpha'' = \alpha_1$. This combination of α' and α'' reproduces the equilibrium of Proposition 2. It thus suffices to show that $\alpha' = 0$ is the only possibility for α'.

Suppose that $\alpha' = \alpha_1$. Then the second-period probability of cooperation on the part of the first-period defector, given on the left in the following inequality, must satisfy

$$\frac{F(\alpha'') - F(\alpha_1)}{1 - F(\alpha_1)} > \alpha_1,$$

in order to ensure the optimality of cooperation for all $\alpha \in [\underline{\alpha}, \alpha_1]$. It is a sufficient condition for this inequality to *fail* for all possible values of $\alpha'' \in [\alpha_1, 1]$ that

$$\frac{F(1) - \alpha^*}{1 - \alpha^*} < \alpha^*,$$

or $F(1) < 2\alpha^* - (\alpha^*)^2$, as stipulated in (11). Hence, (11) ensures that we cannot have $\alpha' = \alpha_1$.

Could we have $\alpha' \in (0, \alpha_1)$? Such an equilibrium requires

$$F(\alpha')/F(\alpha_1) = \alpha'' \tag{19}$$

in order to sustain cooperation on $[\alpha_1, \alpha'']$, and requires (using (19) for the first equality)

$$\frac{F(\alpha'') - F(\alpha_1)}{1 - F(\alpha_1)} = \frac{F(\frac{F(\alpha')}{F(\alpha_1)}) - F(\alpha_1)}{1 - F(\alpha_1)} = \alpha', \tag{20}$$

in order to sustain cooperation on $[\underline{\alpha}, \alpha']$. We have established that (20) cannot hold for $\alpha = 0$ or $\alpha = \alpha_1$:

$$\frac{F(\frac{F(0)}{F(\alpha_1)}) - F(\alpha_1)}{1 - F(\alpha_1)} < 0 \qquad \frac{F(\frac{F(\alpha_1)}{F(\alpha_1)}) - F(\alpha_1)}{1 - F(\alpha_1)} < \alpha_1 \tag{21}$$

If F is linear, then the middle term in (20) is linear in α'. The inequalities in (21) then preclude the satisfaction of (20) for any $\alpha' \in (0, \alpha')$. Condition (20) will thus fail, and hence the equilibrium characterized in Proposition 2 will be unique, as long as F is not too nonlinear, i.e., as long as $|F''|$ is not too large.

This uniqueness continues to hold for small Θ_C and Θ_D. In particular, the heart of the argument given is that zero is the only possible value for α' because the posterior beliefs following an observation of cooperation and an observation of defection are too far apart. As long as Θ_C and Θ_D are small and hence posterior expectations sufficiently close to those of (14), they will remain too far apart to admit any equilibrium other than the counterpart of the noiseless equilibrium. ||

References

[1] James Andreoni, Marco Castillo, and Ragan Petrie. What do bargainers' preferences look like? Exploring a convex ultimatum game. *American Economic Review*, 2003. Forthcoming.

[2] James Andreoni and John H. Miller. Rational cooperation in the finitely repeated prisoners' dilemma: Experimental evidence. *Economic Journal*, 103:570–585, 1993.

[3] James Andreoni and John H. Miller. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70:737–753, 2002.

[4] Kenneth Arrow. Gifts and exchanges. *Philosophy and Public Affairs*, 1:343–362, 1972.

[5] Ken Binmore, John McCarthy, Giovanni Ponti, Larry Samuelson, and Avner Shaked. A backward induction experiment. *Journal of Economic Theory*, 104:48–88, 2002.

[6] Ken Binmore, Chris Proulx, Larry Samuelson, and Joe Swierzbinski. Hard bargains and lost opportunities. *Economic Journal*, 108:1279–1298, 1998.

[7] Matthias Blonski and Daniel A. Probst. The emergence of trust. Mimeo, University of Mannheim, 2001.

[8] Lorne Carmichael and Bentley MacLeod. Gift giving and the evolution of cooperation. *International Economic Review*, 38:485–510, 1997.

[9] Saikat Datta. Building trust. STICERD Discussion Paper 96/305, London School of Economics, 1996.

[10] Douglas W. Diamond. Reputation acquisition in debt markets. *Journal of Political Economy*, 97:828–862, 1989.

[11] Armin Falk, Ernst Fehr, and Urs Fischbacher. On the nature of fair behavior. *Economic Inquiry*, 2002. Forthcoming.

[12] Parikshit Ghosh and Debraj Ray. Cooperation in community interaction without information flows. *Review of Economic Studies*, 63:491–519, 1996.

[13] S. Knack and P. Keefer. Does social capital have an economic payoff? A cross country investigation. *Quarterly Journal of Economics*, 112:1251–1288, 1997.

[14] Robert D. Putnam. *Bowling Alone*. Simon and Schuster, New York, 2000.

[15] Matthew Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83:1281–1302, 1993.

[16] Joel Watson. Starting small and renegotiation. *Journal of Economic Theory*, 85:52–90, 1999.

[17] Joel Watson. Starting small and commitment. *Games and Economic Behavior*, 38:176–199, 2002.

[18] Paul F. Whiteley. Economic growth and social captial. *Political Studies*, 48:443–466, 2000.