Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Economic Theory 127 (2006) 117–154
JOURNAL OF Economic Theory
www.elsevier.com/locate/jet

# Building rational cooperation

James Andreoni, Larry Samuelson*

*Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706-1321, USA*

Received 3 August 2004; final version received 24 September 2004
Available online 17 January 2005

**Abstract**

Experiments have shown that people have a natural taste for cooperation. This paper takes a first step in understanding how formal and informal institutions might be designed to utilize these private tastes to facilitate more efficient economic interactions. We examine a twice-played prisoners' dilemma in which the total of the stakes in the two periods is fixed, but the distribution of these stakes can be varied across periods. We verify experimentally that it is best to "start small," reserving most of the stakes for the second period.


*JEL classification:* C70; C90; D64; Z13

*Keywords:* Cooperation; Prisoners' dilemma; Starting small; Reciprocity

## 1. Introduction

People often behave cooperatively.[1] We rely on such cooperative behavior throughout our daily lives. For example, we routinely complete simple transactions without giving the slightest thought to whether contracts or legal resources will deter the opportunistic behavior that is inevitably possible. At the same time, cooperation is more readily elicited in some interactions than others. One may trust the quality of a watch purchased at a jewelry store but not one purchased on a subway platform.

* Corresponding author. Fax: +1 608 262 2033.
*E-mail addresses:* andreoni@wisc.edu (J. Andreoni), larrysam@ssc.wisc.edu (L. Samuelson).

[1] See Andreoni and Miller [1] or Camerer [7, Chapter 2] for recent discussions of the (vast) relevant experimental literature.

0022-0531/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jet.2004.09.002

This paper takes a first step in asking a basic question for economic analysis: Are there simple ways to structure interactions so as to enhance economic efficiency by fostering cooperation? We are especially interested in the effect of starting relationships with small stakes, allowing people a low-cost opportunity to exhibit a willingness to cooperate and to assess the propensity of others to cooperate, making investments in cooperation that can pay off in the form of mutual cooperation for larger subsequent stakes.[2] An employer forming a new team of workers may give them small initial tasks, to help build cooperation, followed by larger tasks that can take advantage of that cooperation. Wary countries begin with cultural exchanges, work up to diplomatic and economic ties, negotiate military treaties, and finally redirect their military resources elsewhere.

While it seems intuitive that starting small can foster cooperation, the argument 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 that can be exploited at higher stakes, prompting suspicions of initial cooperators. Identifying circumstances conducive to cooperation thus requires theoretical modeling and experimental work. We view our work as a first step toward understanding these issues.

### *1.1. Rational cooperation*

We place our analysis in an extremely simple environment, that of a twice-played prisoners' dilemma. Experiments with the prisoners' dilemma consistently find:[3]

- 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," cooperating 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 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. If we are to retain the unifying theme of economics, that people can be usefully modeled as making rational choices, we must then ask whether we can rationalize the observed behavior by expanding our notion of preferences to include additional considerations.

---

[2] Binmore et al. [4] 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. Theoretical models in which relationships optimally start small are examined by Blonski and Probst [5], Datta [10], Diamond [11] and Watson [22,23].

[3] See, for example, Andreoni and Miller [2], Rabin [20], Roth and Murnighan [21], and their references.

A number of specifications have been suggested for preferences that will induce cooperation in the prisoners' dilemma (e.g., Bolton and Ockenfels [6], Falk and Fischbacher [13], Fehr and Schmidt [14], Levine [17], and Rabin [20]). The difficulty in interpreting such models is distinguishing when we have uncovered a robust feature of behavior and when we have fortuitously constructed preferences that happen to match some experimental observations.

Our confidence in a specification of preferences is enhanced to the extent that the model makes additional predictions that can be tested experimentally. Toward this end, this paper begins, in Section 2, with a model of preferences capturing the features that are common to the various specifications offered in the literature as explanations of cooperation in the prisoners' dilemma. Section 3 derives some predictions of the model concerning behavior in the twice-played prisoners' dilemma, including predictions for the ability to enhance cooperation by starting with relatively small stakes. Sections 4 and 5 report the results of an experiment examining these predictions. Section 5 also discusses limitations of the analysis and considers the relationship of our work to explanations based on departures from rationality. Section 6 concludes.

### *1.2. 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. While the various models that have been offered in the literature differ in many details, these features provide a concise summary of their common ground.[4]

Preferences exhibiting these four properties easily 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 Fig. 1.

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

Period one

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

Period two

Fig. 1. Stage games for the twice-played prisoners' dilemma, where $x_1, x_2 \geqslant 0$.

[4] For example, Rabin [20], citing evidence from psychology for these assumptions, designed a model of fairness to capture them. The one-period version of our model will have many predictions in common with Rabin's, but without specifying the intentions or motives of our players.

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 $\lambda$ ranges from zero-to-one. When $\lambda = 0$, all of the payoffs are concentrated in the first of the two prisoners' dilemmas. As $\lambda$ 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 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, we have the following results:

- 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 $\lambda$ does.
- Certain outcomes of the game (identified below) become more likely, and others less likely, as $\lambda$ grows. For example, when $\lambda$ is small, we predict that an outcome of mutual cooperation in the first period should be followed by mutual cooperation in the second. However, as $\lambda$ increases above a threshold, it becomes increasingly likely that mutual cooperation is followed by one or both players defecting.

The behavioral patterns outlined in the previous point give rise to conflicting effects on payoffs that, with some functional-form assumptions, combine to produce what we regard as an intuitive effect:

- The expected monetary payoff from the two-period game initially increases in $\lambda$, achieves an interior maximum at a value of $\lambda > \frac{1}{2}$, 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 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 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 $\lambda$. As a result, higher values of $\lambda$ induce agents to cooperate more in the first period as an investment in second-period cooperation, as well as inducing a number of more specific behavioral shifts that we describe below. Finally, as $\lambda$ increases, we trade off increased first-period cooperation for decreased first-period payoffs, as payoffs are shifted to the second period. The combined effects suggest that monetary payoffs will be minimized when $\lambda = 0$ or 1, and will achieve an interior maximum. In particular, we find payoffs are maximized when the second-period stakes are one-and-one-half to two times as large as those of the first period.

## 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 utilities from cooperating ($C$) and defecting ($D$) are given by

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

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

where the continuous function

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

identifies an agent's expected utility as a function of the action $z \in \{C, D\}$ chosen by the agent, the probability $\rho$ with which the agent's opponent cooperates,[5] and the value $\alpha$ 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 utility of cooperating (or defecting) is perturbed by a random variable $\Theta_C$ (or $\Theta_D$). The random variables $\Theta_C$ and $\Theta_D$ are independent and have zero means, with distribution functions that have strictly positive densities on the reals (i.e., have full support and no mass points). The realized values $\theta_C$ and $\theta_D$ of these random variables are drawn before the agent makes his choice.

Second, the agent's realized utility depends upon the opponents' action. We think of the primitives of our model as utility functions $\tilde{\pi}(C, C, \alpha)$ and $\tilde{\pi}(C, D, \alpha)$, giving the utility of cooperation when the opponent cooperates and defects, respectively (temporarily ignoring $\Theta_C$). The expected utility, when the opponent cooperates with probability $\rho$, is then $\rho\tilde{\pi}(C, C, \alpha) + (1 - \rho)\tilde{\pi}(C, D, \alpha) \equiv \pi(C, \rho, \alpha)$. The function $\pi(D, \rho, \alpha)$ is similarly constructed. We find it useful to work directly with $\pi(C, \rho, \alpha)$ and $\pi(D, \rho, \alpha)$.

The random variables $\Theta_C$ and $\Theta_D$, 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 $\Theta_C$ and $\Theta_D$ are *sufficiently small*, meaning that the distributions of $\Theta_C$ and $\Theta_D$ 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 $\Theta_C$ and $\Theta_D$ 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 $\Theta_C$ and $\Theta_D$ are similarly homogeneous.[6] Equilibrium play in a one-shot prisoners' dilemma would thus be unaffected by the stakes of the game, in the

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

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

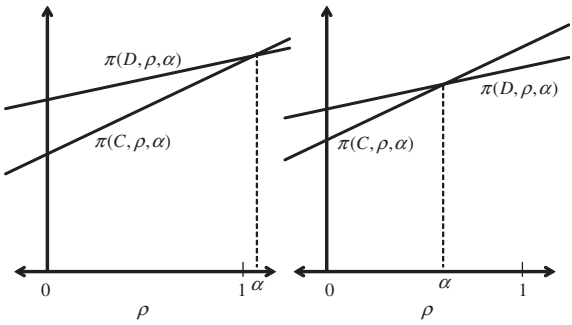


Fig. 2. Possible utilities of cooperation ($\pi(C, \rho, \alpha)$) and defection ($\pi(D, \rho, \alpha)$) as a function of probability $\rho$ that an opponent cooperates (cf. footnote 5).

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, building a notion of reciprocity into preferences:

**Assumption 1.** For all $(z, \rho, \alpha)$,

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

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. Fig. 2 illustrates preferences that are consistent with this assumption. Condition (4) is consistent with standard models of self-interested preferences, while condition (5) is the departure that potentially makes cooperation optimal. If the inequality in condition (5) is reversed, then defection becomes more attractive the more likely 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 $\alpha$ are assigned to players so that a player characterized by $\alpha$ is indifferent between $C$ and $D$, given $\theta_C = \theta_D = 0$, when his opponent plays $C$ with probability $\alpha$, i.e.,

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

Equivalently, $\alpha$ is the probability of opponent cooperation above which a player of characteristic $\alpha$ prefers to cooperate rather than defect. Despite calling $\alpha$ a probability, we do not

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

restrict $\alpha$ to lie within $[0, 1]$. 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 Fig. 2, allowing our model to accommodate agents with "standard" preferences. 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 Fig. 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 $\alpha$. We find it convenient to refer to a player characterized by value $\alpha$ as "player $\alpha$." We think of $\alpha$ as being a characteristic of a player that is fixed, while the perturbations $\theta_C$ and $\theta_D$ are drawn anew each time the game is played.

Assumption 1 is our key behavioral assumption and plays an essential role in the model. The remaining assumptions make the model simple enough to be tractable. We assume:

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

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

This assumption ensures that the parameter $\alpha$ captures all of the information we need about the differing preferences of different agents. Players with different values of $\alpha$ are characterized by utilities $\pi(C, \rho, \alpha)$ and $\pi(D, \rho, \alpha)$ (as a function of $\rho$) that are horizontal shifts of one another.[8]

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

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

$$0 < f(\alpha) < 1. \tag{6}$$

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 of incomplete information. The assumption that $F$ is strictly increasing on $[\underline{\alpha}, \overline{\alpha}]$ ensures that $F(0) > 0$, and $1 - F(1) > 0$, 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

[8] A sufficient condition for this assumption is that an agent maximizes the expected value of a utility function defined over the set $\{C, D\}^2$ (of the agent's and the opponent's choices) as the sum of (i) the resulting monetary payoff, (ii) a payoff that depends upon the pair of actions and is common across agents, and (iii) a payoff that depends upon the agent's own action and is idiosyncratic across agents. Assumption 2 is sufficient for our results but stronger than we need. The key implication we use in our analysis is the single-crossing condition that if second-period play is such that agent $\alpha$ finds it optimal to cooperate in the first period, then so does any "more cooperative" (lower value of $\alpha$) agent.

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

Assumption 3 ensures that there is a unique $\alpha^* \in [\underline{\alpha}, \overline{\alpha}]$ such that $F(\alpha^*) = \alpha^*$. We assume $F$ is such that

**Assumption 4.**

$$F(0) \leqslant (\alpha^*)^2, \tag{7}$$

$$F(1) \leqslant 2\alpha^* - (\alpha^*)^2, \tag{8}$$

$$f(\alpha) < \frac{F(\alpha)}{\alpha} \quad \forall \alpha \in [\alpha^*, 1]. \tag{9}$$

Assumption 4 is designed to ensure that there is sufficient diversity in the distribution of types. The first two conditions in Assumption 4 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.

All of the assumptions we make on $F$ in this paper are satisfied if, for example, $F$ is a uniform distribution on $[-\frac{1}{2}, \frac{3}{2}]$, but this is not necessary.[9]

### 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 $\alpha$ and the realized values of $\theta_C$ and $\theta_D$, but know only that their opponent's values of $\alpha$, $\theta_C$ and $\theta_D$ 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 $\delta$ is the realization of a random variable whose distribution converges weakly to a unitary mass on zero as do the distributions of $\Theta_C$ and $\Theta_D$. Behavior will depend only upon $\delta$ rather than upon the values of $\theta_C$ and $\theta_D$, and we will refer to a player as being characterized by a value of $\alpha$ and a realized value of $\delta$. Section A proves:[10]

[9] More generally, the assumptions hold for any uniform distribution on $[-z, 1 + \phi z]$ for $z, \phi > 0$ with $\frac{\phi}{1+\phi} \leqslant z \leqslant \frac{1}{(1+\phi)\phi}$. Notice that $z = \frac{1}{2}$ is the only case in which the symmetry condition $\phi = 1$ can hold.

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

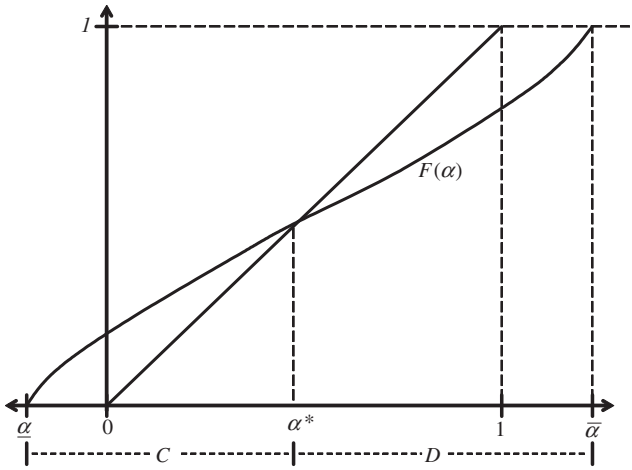


Fig. 3. Equilibrium of one-shot prisoners' dilemma when $\Theta_C$ and $\Theta_D$ are zero.

**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* $(\delta, \alpha)$ *cooperates if* $\delta > \Delta(\alpha)$ *and defects if* $\delta < \Delta(\alpha)$.

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

Hence, players with smaller values of $\alpha$ and larger values of $\delta$ are more likely to cooperate. Fig. 3 illustrates the equilibrium for the limiting case in which the perturbations $\Theta_C$ and $\Theta_D$ are arbitrarily small. The equilibrium value $\alpha^*$ 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.

Notice that a complete-information game between two agents whose values of $\alpha$ are known may (or may not, depending upon the agents' values of $\alpha$) feature multiple equilibria, including one in which both agents cooperate and one in which both defect. The uniqueness established in Proposition 1 pertains to the *incomplete-information* game whose prior is given by $F$.

Consider two populations characterized by distributions $F$ and $F'$ over $\alpha$ with $F$ first-order stochastically dominating $F'$, so players in population $F$ are characterized by higher values of $\alpha$ and hence are less likely to cooperate. We have:[11]

[11] In the notation of the proof of Proposition 1, this follows from the observation that, for any fixed $\tilde{\Delta}(\alpha, \rho)$, $F$ induces a smaller value of $R(\rho)$ than does $F'$.

**Corollary 1.** *Let Assumptions* 1–3 *hold. If distribution* $F$ (*over* $\alpha$) *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* $F$.

*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 $\alpha^*$ 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 2 in the appendix presents an equilibrium in which all agents defect in the first period, because the out-of-equilibrium event of cooperating in the first period prompts beliefs that the agent (and hence the opponent) is *less* likely to cooperate in the second period than would otherwise be the case. We find such equilibria counter-intuitive and suspect they are unlikely to provide good descriptions of the play of experimental subjects, and hence 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 higher when player $i$ cooperates in period one than when player $i$ defects.

We have a convenient characterization of first-period behavior in monotonic equilibria. The appendix proves:[12]

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

**Lemma 2.** *In a monotonic equilibrium,*

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

The first result is straightforward, indicating that first-period cooperation occurs among those agents who are most inclined to cooperate (i.e., have small values of $\alpha$ and large $\delta$). Next, in a monotonic equilibrium, cooperating in the first period makes it more likely that one's opponent will cooperate in the next period. Hence, as Lemma 2 indicates, players are more inclined to cooperate in the first period of a two-period game than in a one-shot game.

The following proposition characterizes the second-period equilibrium of the two-period game. For the limiting case in which $\Theta_C$ and $\Theta_D$ are zero, we have a uniqueness result for the second-period equilibrium that leads to a straightforward existence result for the two-period game. When $\Theta_C$ and $\Theta_D$ are nonzero but small, we cannot exclude the possibility of multiple second-stage equilibria, all of which must satisfy our characterization and all of which must be close to the unique equilibrium of the noiseless limit. A somewhat more complicated argument then establishes existence in the two-period game.

[12] As Table 2 below indicates, our subjects' behavior is consistent with monotonicity.

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

(2.1) *As implied by Lemmas* 1 *and* 2, *let* $\Delta_1(\alpha)$ *be increasing with* $\Delta_1(\alpha) < \Delta^*(\alpha)$ *for all* $\alpha$ *and with player* $(\delta, \alpha)$ *cooperating in the first period if and only if* $\delta > \Delta_1(\alpha)$. *Then equilibrium second-period behavior 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* $(\delta, \alpha)$ *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* $(\delta, \alpha)$ *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* $\Delta_{\{2,CD\}}(\alpha) > \Delta_{\{2,DC\}}(\alpha) > \Delta_1(\alpha)$ *such that player* $i$ $(j)$ *cooperates in period two if and only if* $\delta(i) > \Delta_{\{2,CD\}}(\alpha(i))$ $(\delta(j) > \Delta_{\{2,DC\}}(\alpha(j)))$.

*In the limiting case of* $\Theta_C = \Theta_D = 0$, *this second-period equilibrium is unique.*

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

The intuition behind this result is that having an opponent cooperate in the first period is good news about the opponent's propensity to cooperate, while a defection is bad news. Play in the first period of the two-period game is qualitatively like that of the one-period game, but with the prospect of the second period inducing more cooperation ($\Delta_1(\alpha) < \Delta^*(\alpha)$). If both players cooperate in the first period, both have received good news and play in the second period again looks much the same, but with even more cooperation ($\Delta_{\{2,CC\}}(\alpha) < \Delta_1(\alpha)$). Mutual defection conveys bad news to both players, and puts a damper on second-period cooperation ($\Delta_{\{2,DD\}}(\alpha) > \Delta_1(\alpha)$). If player $i$ cooperates and $j$ defects, then $i$ believes $j$ is a type who is unlikely to cooperate in the second period. Player $i$ then cooperates only if $i$'s value of $\alpha$ is so low as to make cooperation optimal when $j$ is quite likely to defect. Player $j$ has received encouraging news about $i$ in this case, but in light of player $i$'s behavior player $j$ also defects unless her value of $\alpha$ happens to be quite low (which is almost certainly not the case, given that she defected in the first period).

We can illustrate these properties when there is no noise ($\Theta_C$ and $\Theta_D$ are zero):

**Remark 1.** If $\Theta_C$ and $\Theta_D$ are zero, then the first period is characterized by a value $\alpha_1 > \alpha^*$ such that in the first period, player $\alpha$ cooperates if and only if $\alpha < \alpha_1$. Second-period behavior is then given by

1. Let $\alpha_2 > \alpha_1$ be the unique solution to

$$\frac{F(\alpha_2)}{F(\alpha_1)} = \alpha_2 \tag{10}$$

if $\alpha_1 > 1$, and let $\alpha_2 = 1$ if $\alpha_1 < 1$. Hence, $\alpha_2$ is determined just as is $\alpha^*$ in the single-period game, with the prior $F(\alpha)$ replaced by the posterior $F(\alpha)/F(\alpha_1)$, which is relevant for both players if both cooperate in the first period. Fig. 4 illustrates. If both players

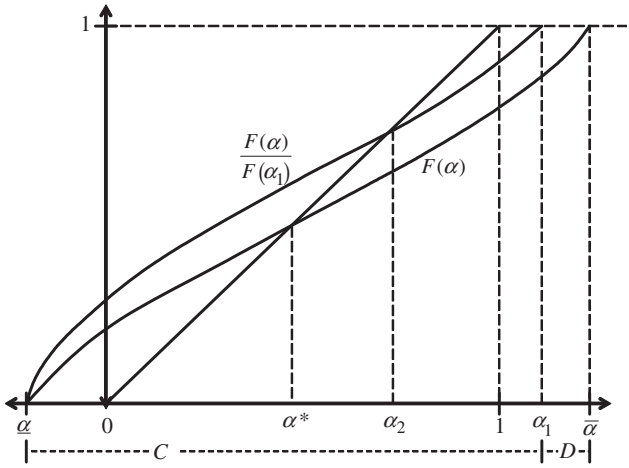


Fig. 4. Possible second-stage equilibrium following $(C, C)$.

cooperate in the first period, then player $\alpha$ cooperates in the second period if and only if $\alpha < \alpha_2$.

2. If both players defect in the first period, then player $\alpha$ 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 $i$ cooperates in the second period if and only if $\alpha(i) < 0$, while player $j$ defects in the second period.

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 $\Delta_1(\alpha)$. 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.[13]

## 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 on the presumptions that utilities are homogeneous of degree one in monetary payoffs and that the perturbations $\Theta_C$ and $\Theta_D$ are small.

Let $\hat{\Delta}_1(\alpha, \lambda)$ be the equilibrium function $\Delta_1(\alpha)$ (cf. Proposition 2) in the maximally cooperative equilibrium given $\lambda$. Let an outcome of the two-period game be written as (for

[13] More precisely, if there are multiple limiting equilibria as $\Theta_C$ and $\Theta_D$ approach 0, then the maximally cooperative one has the largest value of $\alpha_1$, say $\bar{\alpha}_1$. When considering cases where $\Theta_C$ and $\Theta_D$ are nonzero but small, we restrict attention to equilibria that can be taken to be arbitrarily close to the limiting equilibrium characterized by $\bar{\alpha}_1$.

example) $(CD, DD)$, in which case player 1 cooperated and 2 defected in the first period, with both players defecting in the second period. Then we have

**Proposition 3.** *Let Assumptions* 1–4 *hold and let* $\lambda \in (0, 1)$ *unless otherwise stated. Then, for sufficiently small* $\Theta_C$ *and* $\Theta_D$:[14]

(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* $\alpha$, $\Theta_C$ *and* $\Theta_D$ *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*

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

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

(3.4) *The expected incidences of* $(CC, DD)$, $(CC, CD)$ *and* $(CC, DC)$ *are approximately zero for small values of* $\lambda$ (*values of* $\lambda$ *for which* $\alpha_1 < 1$, *when* $\Theta_C = \Theta_D = 0$), *with* $(CC, DD)$ *thereafter increasing in* $\lambda$ *and* $(CC, CD)$ *and* $(CC, DC)$ *thereafter positive but with an ambiguous comparative static in* $\lambda$. *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 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 $\lambda$, but is greater than the incentive to cooperate in the second period for every value $\lambda \in (0, 1)$. Hence, the shadow of the future enhances current cooperation.[15]

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.

[14] We state the following results for the limiting case of $\Theta_C = \Theta_D = 0$. If $\Theta_C$ and $\Theta_D$ are nonzero but small, Proposition 3.1 holds as stated. Proposition 3.2 becomes the statement that play in these two games is (arbitrarily, as $\Theta_C$ and $\Theta_D$ approach zero) close. The corresponding comparative static results in the final two sections are that for any $\lambda$ and $\lambda'$, then the implied inequality in behavior holds if $\Theta_C$ and $\Theta_D$ are sufficiently small.

[15] When $\lambda = 0$ or 1, one period is irrelevant, in which play is arbitrary. 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- and second-period cooperation.

Statement (3.4) provides predictions for those two-period outcomes for which results are unambiguous.[16] The first part indicates that when $\lambda$ is small (so that $\alpha_1 < 1$) and hence the first period relatively important, mutual cooperation in the first period is followed by mutual cooperation in the second period. As $\lambda$ increases (and hence $\alpha_1 > 1$), the increased second-period payoffs give rise to outcomes in which cooperation in the first period sets up second-period defections. In addition, we identify specific outcomes involving first-period defection (($CD, CD$), ($DC, DC$) and ($DD, DD$)) that become less likely as the second period becomes more important.

**Example 1.** The behavior described in Proposition 3.4 gives rise to conflicting effects on overall payoffs (i.e., the sum the two players' payoffs over the two periods) as $\lambda$ increases, since an increase in $\lambda$ induces more cooperation in the first period but transfers payoffs to the relatively defection-ridden second period. The net effect is ambiguous without some additional assumptions about functional forms. We present here an example capturing our intuition that these conflicting forces will combine to cause overall payoffs to first increase and then decrease in $\lambda$, finding their maximum at a value $\lambda > \frac{1}{2}$, when second-period payoffs are sufficiently important but not arbitrarily larger than first-period payoffs. Let us simplify the calculations by taking $\Theta_C = \Theta_D = 0$. As in Example 2 in the appendix, let $\alpha$ be uniform on $[-\frac{1}{2}, \frac{3}{2}]$ and let

$$\pi(C, \rho, \alpha) = \tfrac{5}{8}(\rho - \alpha),$$
$$\pi(D, \rho, \alpha) = \tfrac{1}{8}(\rho - \alpha).$$

Then it is tedious but straightforward to verify that $\alpha_1$ increases from $\frac{1}{2}$ to 1 as $\lambda$ increases from 0 to $\frac{2}{3}$. Overall payoffs increase as $\lambda$ increases from 0 to $\frac{2}{3}$, and then eventually fall, attaining their minimum value at $\lambda = 1$ (duplicating the value attached to $\lambda = 0$). Maximizing the surplus thus calls for the relationship to begin with relatively small stakes and feature larger stakes in the second period.

**Proof of Proposition 3.** We consider the limiting case in which $\Theta_C = \Theta_D = 0$. The results for small but nonzero values of $\Theta_C$ and $\Theta_D$ (cf. footnote 14) follow immediately.

Let

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

be a function with the property that $H(\alpha, \alpha')$ identifies the equilibrium path of play given that the players' actual types are $\alpha$ and $\alpha'$. Hence, $H(\alpha, \alpha') = \{DC, DD\}$ indicates that players $\alpha$ and $\alpha'$ 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

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

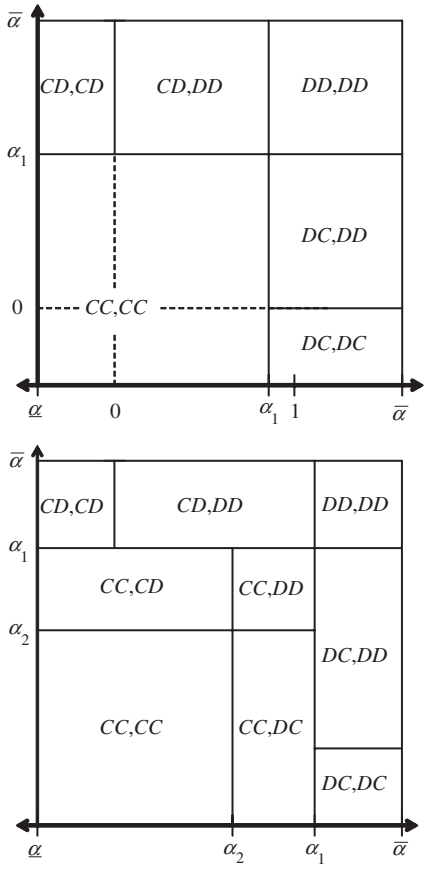


Fig. 5. Function $H(\alpha, \alpha')$ corresponding to a monotonic equilibrium, where $\alpha_1 < 1$ (top) and $\alpha_1 > 1$ (bottom), $(\Theta_C = \Theta_D = 0)$.

agent $\alpha$ when paired with agent $\alpha'$. Then expected equilibrium payoffs are given by

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

Fig. 5 illustrates the function $H(\alpha, \alpha')$ for an equilibrium in which $\alpha_1 < 1$ and an equilibrium in which $\alpha_1 > 1$.

The results now follow from examining the function $H(\alpha, \alpha')$. In particular, for small values of $\lambda$, the top panel in Fig. 5 is relevant. We show that as $\lambda$ increases, and hence the second period becomes more important, the value of $\alpha_1$ increases. This continues until $\alpha_1$ hits 1, at which point the bottom panel of Fig. 5 is relevant, initially with $\alpha_2 = \alpha_1$. Subsequent increases in $\lambda$ increase $\alpha_1$ and decrease $\alpha_2$. These movements have implications for the likelihood of the various outcomes shown in Fig. 5, which lead to the results reported in Proposition 3.

(3.1) Fig. 5 shows that when $\Theta_C = \Theta_D = 0$, then for each possible outcome path in a monotonic equilibrium, there is weakly more cooperation in the first period than the second, with a strict inequality for some values of $(\alpha, \alpha')$. Taking an expectation over $(\alpha, \alpha')$, 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 equilibrium outcomes. 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, consider player $\alpha_1(\lambda)$, defined to be the player who is indifferent between $C$ and $D$ in the first period of a maximally cooperative equilibrium, given $\lambda$. Lemma 2 implies that

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

Using the assumption that $\pi$ is linearly homogeneous in monetary payoffs, an increase from $\lambda$ to $\lambda'$ then causes this player to strictly prefer $C$ in the first period. Hence, there exists a set $[\alpha_1(\lambda), \hat{\alpha}] \subset [\alpha_1(\lambda), \overline{\alpha}]$ with the property that, for any $\alpha_1 \in [\alpha_1(\lambda), \hat{\alpha}]$ (and given $\lambda'$ and the unique second-period equilibrium described by Proposition 2, and fixing first-period behavior so that players $\alpha < \alpha_1$ cooperate and $\alpha > \alpha_1$ defect), player $\alpha_1$ at least weakly prefers $C$ in the first period. 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 coupling the first-period behavior of cooperating if and only if $\alpha < \hat{\alpha}$ with the corresponding second-period equilibrium gives an equilibrium (given $\lambda'$) featuring more cooperation than the maximally cooperative equilibrium given $\lambda$ (because $\hat{\alpha} > \alpha_1(\lambda)$), establishing the result. Second, if $\hat{\alpha} = \overline{\alpha}$, then there is an equilibrium with $\alpha_1(\lambda') = \overline{\alpha} > \alpha_1(\lambda)$, again establishing the result.

(3.4) Using the top panel in Fig. 5 as a guide, consider an increase in $\lambda$ and hence $\alpha_1$ (from Proposition 3.3), beginning with the minimum value of $\alpha_1(0) = \alpha^* < 1$ (the latter because $F(1) < 1$). As $\alpha_1$ increases, we have the following possible transitions in behavior, each applicable to some values of $(\alpha, \alpha')$:

$$\begin{aligned}(CD, CD) &\to (CC, CC)\\ (CD, DD) &\to (CC, CC)\\ (DC, DC) &\to (CC, CC)\\ (DC, DD) &\to (CC, CC)\\ (DD, DD) &\to (CC, CC)\\ (DD, DD) &\to (CD, DD)\\ (DD, DD) &\to (DC, DD).\end{aligned} \tag{11}$$

Hence, the incidence of $(CD, CD)$, $(DC, DC)$ and $(DD, DD)$ falls, while $(CC, CC)$ increases. Section A extends the argument to the case in which $\alpha_1$ reaches 1. □

## 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 (22 per session) 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 were known to the subjects.[17]

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 Fig. 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 $\lambda$ drawn from the set $\{0, 0.1, 0.2, \ldots, 1\}$. Table 1 shows the distribution of realized values of $\lambda$ 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 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.[18]

## 5. Results

This section provides a summary of the experimental outcomes and then examines each of the four parts of Proposition 3 as well as Example 1. Our working hypothesis is that each subject is characterized by a realized value of $\alpha$ that remains fixed throughout the experimental session, where these values are independently drawn from a

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

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

Table 1
Values of $\lambda$, with the corresponding period-1 pull value $x_1$ (cf. Fig. 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

| | Period-1 | | Period-2 | | |
|---|---|---|---|---|---|
| $\lambda$ | Pull value | Push value | Pull value | Push value | Frequency |
| 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 |

distribution $F$. We assume that in each period of each two-period game, each player (independently) draws realizations $\theta_C$ and $\theta_D$ of the random variables $\Theta_C$ and $\Theta_D$. 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 20 games. Fig. 6 reports the distribution across subjects of the overall incidence of cooperation.

Fig. 6 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 Eq. (7) 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, Eq. (8) of Assumption 4 may also be reasonable.[19]

Table 2 identifies the outcomes of the 1100 two-period games. Given that each player has two choices in each of two periods, there are 16 possible paths of play. However, we are not interested in distinguishing outcomes that are identical except for which player is labeled

[19] Even committed defectors can optimally cooperate in the first period of the two-period game, so that (8) does not require the observation of agents who always defect, even without appealing to the perturbations $\Theta_C$ and $\Theta_D$.

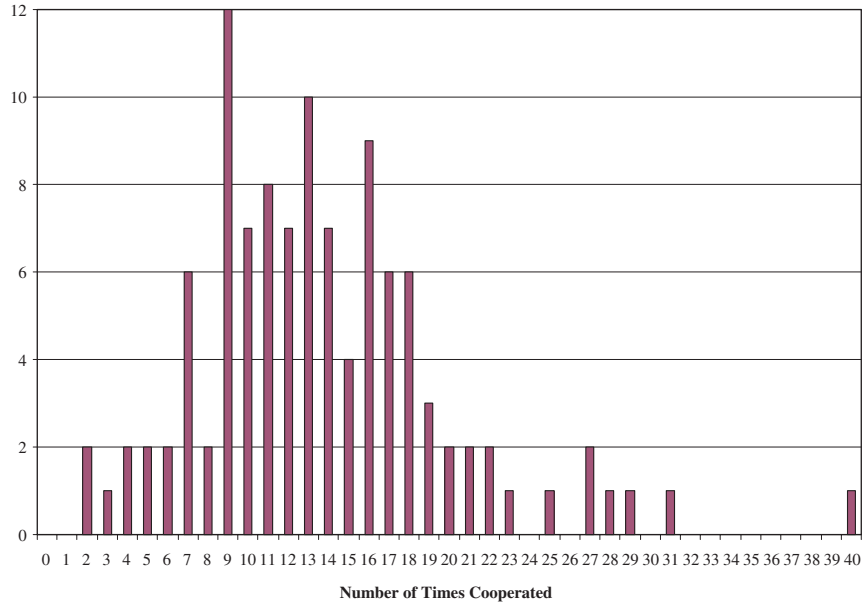


Fig. 6. Each subject faced 40 cooperate/defect decisions. The histogram identifies the number of subjects (vertical axis) exhibiting each of these possible frequencies of cooperation (0–40), on the horizontal axis.

Table 2

| Outcome | | | Frequency | |
|---|---|---|---|---|
| Period one | Period two | | All $\lambda$ | $\lambda \in (0, 1)$ |
| $CC$ | $DD$ | | 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 |
| | | Total | 1100 | 894 |

*Note*. 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.

"player 1" and which "player 2," allowing us to collapse these outcomes to 10 cases. Hence, the last line of Table 2 corresponds to a case in which one player defected in period one (whom we have designated player 1) and one cooperated, while both cooperated in period

two.[20] Table 2 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.[21]

### 5.2. Proposition 3.1: cooperation by period

We begin our assessment of the results with Proposition 3.1, asserting that cooperation will be more prevalent in the first period when $\lambda \in (0, 1)$. From Table 2, we have

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

Cooperation is thus about twice as likely in the first period.

We can disaggregate these data by subject and assess the significance of the difference. Let $\eta_{i1}$ and $\eta_{i2}$ be the number of times subject $i$ cooperated in periods 1 and 2, and 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 $s_\eta$ be the mean and standard deviation of the sample values $\eta_i$, $i = 1, \ldots, 110$. We can assume the distribution of $z = \overline{\eta}/(s_\eta/\sqrt{110})$ 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).[22] We find $\overline{\eta} = 4.33$, $s_n = 4.18$, and $z = 10.9$ ($p \approx 0.000$, with the null hypothesis of no greater cooperation in period two), indicating significantly more cooperation in the first period.

At the same time, this effect does not appear uniformly across values of $\lambda$. The percentage of first- and second-period choices in which subjects cooperated, by value of $\lambda$, is given in Table 3 . Cooperation is more likely in the second period of those games ($\lambda = 0.1$, 0.2, and 0.3) in which the first period is relatively important and second-period stakes are particularly small. We return to this issue in Section 5.7.

### 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

[20] To construct Table 2, 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 one and only one defected in period two, 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.

[21] Lines 5, 6, 9, and 10 of Table 2 correspond to outcomes not predicted by a noiseless version of the model. We comment on these observations in Section 5.7.

[22] The value $\eta_{ij}$ is the result of 20 Bernoulli trials whose mean is unknown and idiosyncratic to player $i$, depending upon $i$'s draw of the parameter $\alpha$. The draws determining $\eta_{i1}$ are independent, but $\eta_{i2}$ is not independent of $\eta_{i1}$.

Table 3
Percentage of cooperation in period one and period two, by value of $\lambda$

| $\lambda$ | Period one | Period two |
|---|---|---|
| 0.1 | 17 | 38 |
| 0.2 | 16 | 27 |
| 0.3 | 16 | 22 |
| 0.4 | 24 | 22 |
| 0.5 | 42 | 14 |
| 0.6 | 55 | 16 |
| 0.7 | 67 | 13 |
| 0.8 | 76 | 14 |
| 0.9 | 78 | 8 |

Table 4

| | Period one $\lambda = 0$ | Period two $\lambda = 1$ |
|---|---|---|
| Number of cooperators | 27 | 22 |
| Percentage | 13.4 | 10.5 |
| Number of defectors | 175 | 188 |

*Note*. 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.

concentrated in the other game. They differ in whether the zero-payoff period comes first ($\lambda = 1$) or second ($\lambda = 0$).

Table 4 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, but not identical.

We can check the statistical significance of the difference by letting $\eta_{i1}$ now be the percentage of the time that player $i$ cooperated when playing in period one of a game with $\lambda = 0$ and letting $\eta_{i2}$ be the percentage of the time that player $i$ cooperated when playing in period two of a game with $\lambda = 1$.[23] 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 $\eta_i = \eta_{i1} - \eta_{i2}$, for those 80 subjects who faced games with both $\lambda = 0$ and 1. Letting $\overline{\eta}$ and $s_\eta$ be the mean and standard deviation of the observed values of $\eta_i$, the distribution of $z = \overline{\eta}/(s_\eta/\sqrt{80})$ is approximately Normal with zero mean and unitary variance. Then we calculate $\eta = 2.3$, $s_n = 28.3$ and $z = 0.71$ ($p = 0.48$, with a null hypothesis

[23] We now work with percentages because a given subject played different numbers of $\lambda = 0$ and 1 games. 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 of 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 1.) These data can be included, at the cost of making an assumption concerning homogeneous 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 in this sample.

Table 5
Results of logit regression

| Independent variable | All rounds | Last 10 rounds |
|---|---|---|
| $\lambda$ | 6.1 | 9.4 |
| | (0.31; $p \approx 0.000$) | (0.78; $p \approx 0.000$) |
| $n$ | 1990 | 995 |
| Round and individual fixed effects | Yes | Yes |

*Note.* There is one observation for each choice by each agent in the first period of each game played by the agent in which $\lambda \in [0, 1)$. The dependent variable equals 1 if the subject cooperated in period one and equals zero otherwise. The first line reports the estimated coefficient on $\lambda$ and the corresponding standard error and $p$-value.

of no difference between periods). The difference in cooperation between periods is thus not significant in this sample.

Is the difference we observe between periods important? One way of assessing this is to ask how the difference in cooperation rates across these two periods compares to either the differences in first- and second-period cooperation reported in Section 5.2 or to the different rates of first-period cooperation induced by different values of $\lambda$ reported in Section 5.4 Each of the latter differences dwarfs the former. Hence, play in those periods which are strategically equivalent (because they are the relevant period an interaction with only one relevant period, whether the first or second) is much more similar than play in periods that are strategically different. While clearly not perfect, the subjects appear able to react to strategically relevant features of the game and ignore irrelevant ones.

### *5.4. Proposition 3.3: first-period cooperation*

Proposition 3.3 predicts that the incidence of first-period cooperation should be increasing in $\lambda$ when $\lambda \in [0, 1)$. The higher is $\lambda$, 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. Table 3 reports first-period cooperation as a function of $\lambda$ for values of $\lambda \in (0, 1)$, while Table 4 shows that 13% of subjects cooperated in the first period when $\lambda = 0$. The incidence of cooperation thus increases from 13% when $\lambda = 0$ to 78% when $\lambda = 0.9$.

To confirm this seemingly obvious link between the concentration of payoffs in the second period and first-period cooperation, Table 5 presents the results of a logit regression in which the dependent variable equals one if a player cooperated in period one and 0 otherwise. Independent variables include $\lambda$, a constant, and dummy variables to identify the round of the observation.[24] In addition, our model 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,

[24] Subjects tend to be somewhat less cooperative in later rounds, but respond to $\lambda$ consistently throughout the experiment. We return to this issue in Section 5.7.

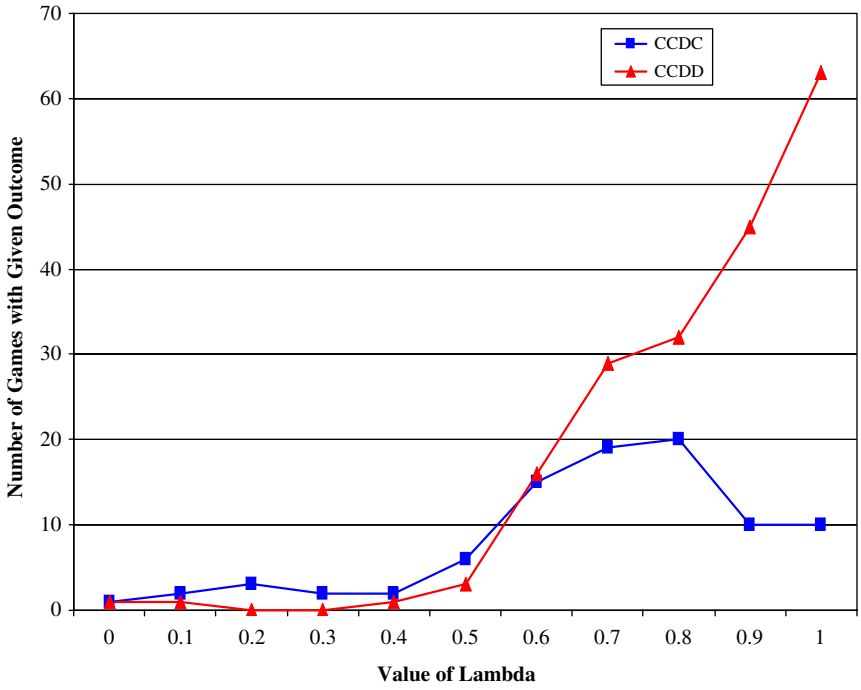


Fig. 7. Frequency of outcomes $(CC, DD)$, and $(CC, DC)$ as a function of $\lambda$. Proposition 3.4 predicts the incidence of $(CC, DD)$ and $(CC, DC)$ will be approximately zero over a common range of relatively small values of $\lambda$, above which $(CC, DD)$ is increasing and $(CC, DC)$ is positive but with an ambiguous comparative static.

we include fixed effects for the players in the regression.[25] We include results for all of the data and for data from the last ten rounds only, returning to differences in the results in Section 5.7. As expected, the estimated coefficient on $\lambda$ is significantly positive: higher period-two payoffs yield higher first-period cooperation.

### 5.5. *Proposition 3.4: paths of play*

Proposition 3.4 makes 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 $\theta_C$ and $\theta_D$ are zero) for relatively small values of $\lambda$, with $(CC, DD)$ increasing

[25] 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 encounters $i$. (Recall that no two players ever meet more than once.)

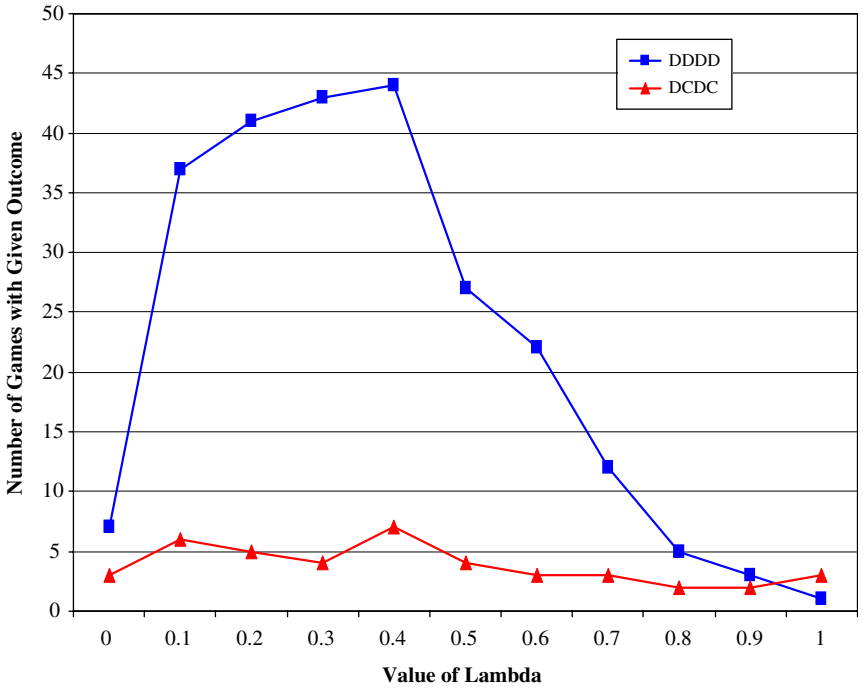


Fig. 8. Frequency of outcomes $(DD, DD)$ and $(DC, DC)$ as a function of $\lambda$. Proposition 3.4 makes no prediction for $\lambda = 0$, and predicts that both will fall as $\lambda$ increases for those values of $\lambda \in (0, 1)$.

for larger values of $\lambda$ and with $(CC, DC)$ positive but with an ambiguous comparative static for these larger values.[26] Fig. 7 plots the incidence of these outcomes as a function of $\lambda$. The predicted trends are visible. Moreover, the common value of $\lambda$ above which these outcomes appear more frequently, corresponding to $\alpha_1 = 1$, is expected to be below the value of $\lambda$ that maximizes the sum of the two-players' expected payoffs, which is consistent with the findings reported below. Hence, when the first period is relatively important, we do not observe first-period mutual cooperation followed by subsequent defection. When the second period is more important, mutual cooperation in the first period is commonly exploited by subsequent defection. These results provide perhaps the most striking support for the model.

The incidences of $(DC, DC)$ and $(DD, DD)$ are predicted to decrease in $\lambda$ for $\lambda \in (0, 1)$. Fig. 8 shows the relevant results. Small sample sizes obscure a hint of a downward trend in the incidence of $(DC, DC)$ as $\lambda$ rises (note the difference in vertical scales in Figs. 7 and 8). 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

[26] Note that the outcomes $(CC, CD)$ and $(CC, DC)$ that appear in Proposition 3.4 are observationally equivalent (cf. Fig. 2), as are the outcomes $(CD, CD)$ and $(DC, DC)$, discussed below.

Table 6
Average value of total payoffs for each value of $\lambda$

| Value of $\lambda$ | 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 |

*Note.* "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 $\lambda$ is given in Table 1.

often for higher values of $\lambda$, 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.

### 5.6. Total payoffs

Example 1 captures our intuition that total payoffs from the two-period game should be first increasing in $\lambda$ and then decreasing in $\lambda$, attaining an interior maximum. The cases of $\lambda = 0$ and 1 should give equal payoffs, reflecting the equivalent behavior for $\lambda = 0$ and 1. Table 6 shows the average earnings, summed over players and periods, for each value of $\lambda$. 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 1 are quite similar.

To assess the relationship between $\lambda$ and total payoffs, Table 7 reports the results of 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 prisoners' dilemma. Independent variables include $\lambda$, $\lambda^2$, and $\lambda^3$ as well as a constant, individual fixed effects and dummy variables identifying the round of the experiment in which the game is played.[27] There were 1100 observations, one for each of the 11 two-period games played in each of 20 rounds in each of 5 experimental sessions. We are interested in the estimated coefficients

[27] 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 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 1 only if it forces the maximum to occur at $\lambda = \frac{1}{2}$. Investigating the value of the $\lambda$ that maximizes total payoffs thus requires a cubic equation.

Table 7
Total-payoff regression results

| Independent variable | All rounds | Last 10 rounds |
|---|---|---|
| $\lambda$ | −4.5 | −5.5 |
| | (6.7; $p = 0.50$) | (7.3; $p = 0.45$) |
| $\lambda^2$ | 51 | 48 |
| | (16; $p = 0.001$) | (17; $p = 0.001$) |
| $\lambda^3$ | −49 | −53 |
| | (10; $p \approx 0.000$) | (11; $p \approx 0.000$) |
| $n$ | 1100 | 550 |
| Payoff-maximizing | | |
| $\lambda$ | 0.65 | 0.68 |

*Note.* There is one observation for each two-period game. The independent variable is the sum of payoffs earned by the two players over the course of the two periods. The first three lines report the estimated coefficients and the corresponding standard errors and $p$-values for powers of $\lambda$. The regression also includes a constant and fixed effects for individuals and rounds.

on $\lambda$, $\lambda^2$ and $\lambda^3$. The coefficient on $\lambda$ is negative, relatively small, and insignificant. The coefficients on $\lambda^2$ and $\lambda^3$ 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 $\lambda = 0.65$, and then decreases.

Expected payoffs are thus nonmonotonic in $\lambda$. 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.

### 5.7. Missing pieces

While the data are reasonably consistent with the comparative static predictions of the model, it is no surprise that our model does not match the data exactly. Our list of topics for further consideration begins with the points of tension.

#### 5.7.1. Learning

It is commonly suggested that experimental subjects require experience in order to learn how to play a game. Our regressions have allowed some insight into this possibility by including round dummies. These dummies suggest that play in later rounds of the experiment

differs from play in early rounds, with the incidence of cooperation falling across rounds.[28] Summary statistics are consistent with this observation. For example, 34% of the subjects' choices were to cooperate over the course of the entire experiment, but the corresponding figure for choices made in the final ten rounds of play is 26%. The question then arises as to whether our results survive this adjustment, or whether they are an artifact of early, possibly confused, play.

In response, we reproduce our key results with attention restricted to the final ten rounds of play. If we restrict the examination of cooperation by period in Section 5.2 to the final ten rounds of play, we find a premium on first-period cooperation that is, if anything, stronger than that appearing over the course of all twenty rounds:

Cooperative plays in period 1: 35% (306 of 886),
Cooperative plays in period 2: 11% (98 of 886).

Restricting Section 5.4's examination of first-period cooperation to the final ten rounds gives the final column of results in Table 5. Once again, first-period cooperation becomes more prevalent as second-period payoffs become more important, with the effect more pronounced than for the case of all twenty rounds. Restricting Section 5.6's examination of total payoffs as a function of $\lambda$ to the final 10 rounds gives the final column of Table 7. We again find a relationship in which total payoffs first increase and subsequently decrease in $\lambda$, with the estimated maximum now at $\lambda = 0.68$.

The forces captured by our model thus appear alongside some apparent learning behavior. However, the behavior captured by our model persists (with a slight hint of intensifying) over the course of this learning. Our analysis has thus apparently captured some important features of behavior, though future models might usefully incorporate provision for adaptive play.

*5.7.2. The cost of cooperation*

Section 5.2 reports that subjects are more likely to cooperate in the first period than the second, but that this effect does not appear uniformly across values of $\lambda$. The first-round cooperation edge is especially high when $\lambda$ is large (the second period is particularly important), becomes smaller for intermediate values of $\lambda$, and turns negative for small values of $\lambda$ (where the first period is particularly important). This suggests that subjects cooperate more readily when payoffs are small. Fig. 9 illustrates this result by 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 0.4, we compare the first-period play for $\lambda = 0.6$ (40% of the stakes in the first period) with second-period play for $\lambda = 0.4$ (40% of the stakes in the second period). We see that cooperation is more likely in the

[28] For Section 5.4's logit regression analysis of cooperation by value of $\lambda$, the default round was the first, with the dummies attached to rounds four through twenty being significantly negative and increasing slightly in absolute value in later rounds. The value of the dummies in later rounds is approximately $-3$, indicating an effect about half as large as that of $\lambda$. For Section 5.6's regression analysis of total payoff by value of $\lambda$, the default round was the first, with the dummies attached to each subsequent round significantly negative and roughly increasing in absolute value across rounds, with total earnings in later periods about 15 points (out of a maximum of sixty) lower than those in initial rounds, an effect about a third as large as the effect of $\lambda^2$ and $\lambda^3$.

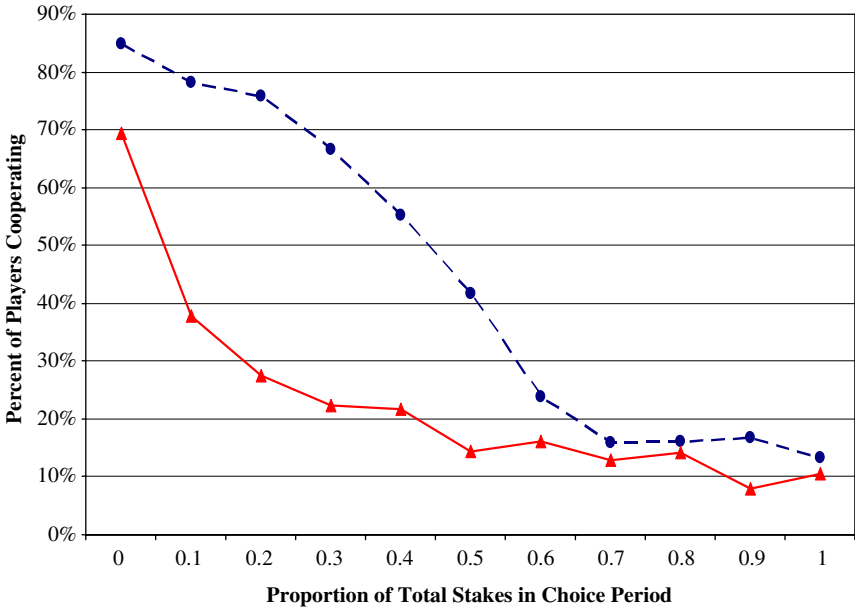


Fig. 9. Incidence of first-period (dashed line) cooperation and second-period (solid line) cooperation, each as a function of the proportion of the total stakes in that period. This groups together first-period choices and second-period choices made with the same stakes in the choice period.

first period than in a similarly sized second period, and that cooperation is more likely in periods with small stakes. This tendency appears in other forms. For example, of the 19 cases in column four of Table 2 in which mutual defection in period one is followed by mutual cooperation in period two ($DDCC$), 16 involve values of $\lambda$ equal to 0.1 of 0.2, and hence relative unimportant second periods. The same is true for 58 of the 90 instances of $DDDC$. Hence, in addition to the forces captured by our model, our subjects appear to be more likely to be cooperative when there is less on the line, again suggesting an interesting extension of the model.

*5.7.3. A comparison*

We have examined a model in which preferences are specified so as to make it rational to sometimes cooperate in the finitely-repeated prisoners' dilemma. An alternative approach, introduced theoretically by the "gang of four" (Kreps, Milgrom, Roberts and Wilson [15,16,19]) and experimentally by Andreoni and Miller [2], Camerer et al. [8], Camerer and Weigelt [9] and McKelvey and Palfrey [18], assumes that most players are rational and have defect as a dominant strategy, but that there is some possibility that one's opponent is an "irrational" or "altruistic" type who sometimes cooperates. This approach is especially interesting in its finding that even small proportions of irrational agents can have large effects on rational behavior.

Our model can be viewed as an extension of such models. A standard "gang-of-four" model would include two types of agents, committed defectors (rational types) and

"irrational" types who play TIT-FOR-TAT. Our model supplements these types with a continuum of irrational types. What do we gain by such an extensions? In a two-period version of the gang-of-four model, we have two possibilities. For small values of $\lambda$, the only sequential equilibrium calls for every rational agent to defect at every opportunity, with variations in $\lambda$ or the period having no effect on behavior. For larger values of $\lambda$, rational agents will all cooperate in the first period and defect in the second period, with variations in $\lambda$ again having no effect.[29] Hence, we should either see universal defection in the first period or universal defection in the second, with the probability of cooperation in the other period in each case equaling the prior probability of a TIT-FOR-TAT agent.

## 6. Discussion

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 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 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. These features are familiar components of rationality-based models of behavior in the prisoners' dilemma behavior. Our analysis is based on implications for how behavior should be affected by changing the relative magnitudes of payoffs across the periods of a twice-played prisoners' dilemma, including the increase in first-period cooperation as second period payoffs become relatively important and the interior maximum of expected total payoffs as a function of the relative importance of second-period payoffs. These predictions appear in the data, leading us to believe that our model captures some robust features of behavior.

There are many more avenues to explore. We are interested in examining games where the relative stakes are chosen by the players. This complicates the analysis but adds a potentially valuable dimension to the game. We also think it important to extend our analysis beyond the two-period prisoners' dilemma game considered here.[30] For example, we would be interested in whether starting small has more of an effect in longer games, in the sense that the optimal ratio between the largest stakes in the game and the initial stakes is higher than it is here. In addition, it would be revealing to apply similar ideas to the "trust game" [3]. We again suspect that trust can be nurtured in repeated play of this game by appropriately choosing the distribution of the stakes.

[29] To see this, note that rational agents must defect in the final period of the model. A rational agent can optimally cooperate in the first period of play only in order to induce TIT-FOR-TAT opponents to also cooperate in the final period. This will be optimal only if $\lambda$ is sufficiently large (relative to the prior probability of a TIT-FOR-TAT opponent).

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

Overall, the results presented here motivate a clear, albeit challenging, interest in exploring how institutions and conventions can be structured to take advantage of tastes for cooperation.

## Acknowledgments

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.

## Appendix A.

**Proof of Proposition 1.** 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 $(\delta, \alpha)$ 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 $\rho$ such that

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

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

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

where $g$ is the density of $\delta$ and $h(\delta, \rho)$ identifies that value of $\alpha$ for which an agent with preference shock $\delta$ is indifferent between cooperating and defecting given probability $\rho$ 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 root, and hence that there is a unique equilibrium.

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

**Example A.1.** 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 $\Theta_C$ and $\Theta_D$ 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 $\alpha$ 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 for all $\alpha$, a perfect Bayesian equilibrium allows the inference that the opponent's value of $\alpha$ 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 $\alpha$ is less than zero, and hence the second-period behavior conditional on a first-period choice of $C$ by either player is again optimal.

We must now 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

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

where the left-hand side is the first-period gain from switching to cooperation and the right-hand 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

$[-\frac{1}{2}, \frac{3}{2}]$ and let preferences be given by

$$\pi(C, \rho, \alpha) = \tfrac{5}{8}(\rho - \alpha),$$
$$\pi(D, \rho, \alpha) = \tfrac{1}{8}(\rho - \alpha).$$

Then $\alpha^* = \frac{1}{2}$ and $F(0) = \frac{1}{4}$, and the required inequality is, for $\alpha < 0$,

$$-\tfrac{5}{8}\alpha - (-\tfrac{1}{8}\alpha) \leqslant \tfrac{5}{8}(\tfrac{1}{2} - \alpha) - \tfrac{5}{8}(\tfrac{1}{4} - \alpha)$$

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

**Proof of Lemma 1.** Consider an agent $\alpha$ and value $\delta$ that prompts cooperation in the first period. Let $\rho_1$ 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 $\alpha$ 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 $\alpha$'s choice requires

$$\pi(C, \rho_1, \alpha) + V(C, \alpha) + \delta \geqslant \pi(D, \rho_1, \alpha) + V(D, \alpha). \tag{12}$$

Letting $\Delta_1(\alpha)$ be the value of $\delta$ that satisfies this relationship with equality, player $\alpha$ will cooperate in period one iff $\delta > \Delta_1(\alpha)$. Now let $\alpha' < \alpha$. Assumptions (1) and (2) imply that, for a monotonic equilibrium,

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

Using this inequality and Assumptions (1) and (2) again, (12) implies

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

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

**Proof of Lemma 2.** Fix a monotonic equilibrium of the two-period game, and let $V_C^*$ and $V_D^*$ be the corresponding equilibrium values of $V(C, \alpha)$ and $V(D, \alpha)$ (introduced in the proof of Lemma 1). Then the functions $\Delta^*(\alpha)$ and $\Delta_1(\alpha)$, respectively, describing behavior in the equilibrium of the single-period game and the first period of the two-period game, must satisfy:

$$\Delta^*(\alpha) = \pi(D, \rho^*, \alpha) - \pi(C, \rho^*, \alpha), \tag{13}$$
$$\Delta_1(\alpha) = \pi(D, \rho_1, \alpha) + V_D^* - \pi(C, \rho_1, \alpha) - V_C^*, \tag{14}$$

where $\rho^*$ and $\rho_1$ are the equilibrium first-period probabilities that the opponent cooperates. Define $\tilde{\Delta}_1(\alpha, \rho)$ to satisfy

$$\pi(C, \rho, \alpha) - \pi(D, \rho, \alpha) + V_C^* - V_D^* + \tilde{\Delta}_1(\alpha, \rho) = 0$$

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 $\rho^*$. From (13)–(14) and the fact that, in a monotonic equilibrium $V_C^* > V_D^*$ (since current cooperation increases the likelihood that the opponent cooperates in the next period), we then have $R_1(\rho^*) > R(\rho^*) = \rho^*$, and hence $R_1(\rho^*) > \rho^*$. Once again, $dR_1(\rho)/d\rho < 1$, and hence the equilibrium value of $\rho_1$ must satisfy $\rho_1 > \rho^*$. Achieving such a higher incidence of first-period cooperation requires $\Delta_1(\alpha) < \Delta^*(\alpha)$. □

**Proof of Proposition 2.** *Step* 1 (Second-period equilibrium): We first note that if $\Theta_C = \Theta_D = 0$, then second-period posterior beliefs about an opponent's type are given by the distribution functions

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

$$\begin{array}{ll} 0 & \text{on } [\underline{\alpha}, \alpha_1] \\ \frac{F(\alpha)-F(\alpha_1)}{1-F(\alpha_1)} & \text{on } [\alpha_1, \overline{\alpha}] \end{array} \quad \text{if } D \text{ observed.} \tag{15}$$

Fig. 10 illustrates these posteriors. Notice that the first distribution is first-order stochastically dominated by $F(\alpha)$, while the second dominates $F(\alpha)$. If $\Theta_C$ and $\Theta_D$ 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 (15) as $\Theta_C$ and $\Theta_D$ get small.

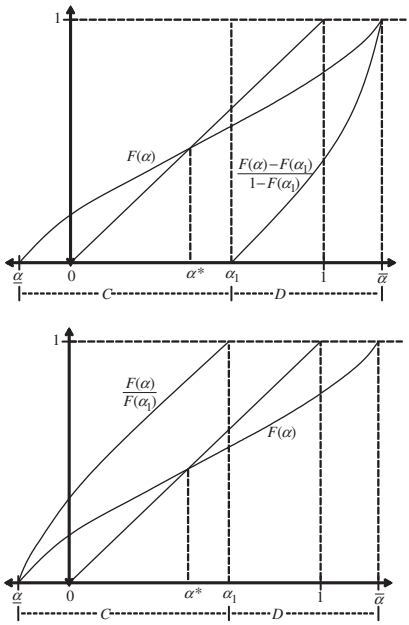


Fig. 10. Posterior beliefs in equilibrium of twice-played prisoners' dilemma, for the case of $\Theta_C$ and $\Theta_D$ equal to zero, given an observation of $D$ (top) or $C$ (bottom, assuming $\alpha_1 < 1$).

We now verify that the second-period strategies constitute an equilibrium. If both agents cooperate in the first period, then the agents enter the second period with symmetric posterior beliefs, given by the first distribution in (15) and illustrated by either by the bottom panel of Fig. 10 (if $\alpha_1 < 1$) or Fig. 4 (in Section 2.3). We can then apply Proposition 1 to ensure the existence and uniqueness of a second-period equilibrium, with some modifications required for the uniqueness argument that are noted in the following step. This equilibrium is described by a function $\Delta_{CC}(\alpha)$, with a player cooperating if and only if $\delta > \Delta_{CC}(\alpha)$. From Corollary 1, we have $\Delta_{CC}(\alpha) < \Delta_1(\alpha)$. Fig. 4 illustrates such a second-period equilibrium, with players cooperating if $\alpha < \alpha_2$ and defecting if $\alpha > \alpha_2$.

If both agents defect in the first period, then the agents again enter the second period with symmetric posterior beliefs, given by the second distribution in (15) and illustrated by the top panel of Fig. 10. We can then apply Proposition 1 to ensure the existence of a second-period equilibrium. This equilibrium is described by a function $\Delta_{DD}(\alpha)$, with a player cooperating if and only if $\delta > \Delta_{DD}(\alpha)$. From Corollary 1, we have $\Delta_{DD}(\alpha) > \Delta_1(\alpha)$. As we approach the limiting case of $\Theta_C = \Theta_D = 0$, it follows from the fact that $\alpha_1$ exceeds the unique value $\alpha^*$ at which $F(\alpha^*) = \alpha^*$ that the function $\Delta_{DD}(\alpha)$ converges to a unique limit featuring $\Delta_{DD}(0) = 0$. Hence, agents $\alpha < 0$ cooperate while agents $\alpha > 0$ defect.

If $\Theta_C = \Theta_D = 0$ and 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 with probability $F(0)/F(\alpha_1)$, making it optimal to defect in the second period for all $\alpha > \alpha_1$ if $F(0)/F(\alpha_1) \leqslant \alpha_1$. It suffices for this inequality that $F(0)/F(\alpha^*) < \alpha^*$ which, using $F(\alpha^*) = \alpha^*$, is ensured by (7). To address the case of nonzero $\Theta_C$ and $\Theta_D$, it is again immediate that following an outcome of $CD$, second-period behavior is described by a pair of increasing functions $\Delta_{\{2,CD\}}(\alpha)$ and $\Delta_{\{2,DC\}}(\alpha)$, with the cooperator (defector) cooperating in the second period if and only if $\delta > \Delta_{\{2,CD\}}(\alpha)$ ($\delta > \Delta_{\{2,DC\}}(\alpha)$). We need to show $\Delta_{\{2,CD\}}(\alpha) > \Delta_{\{2,DC\}}(\alpha) > \Delta_1(\alpha)$. 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 $\Delta_{\{2,CD\}}(\alpha(CD)) = 0 = \Delta_{\{2,DC\}}(\alpha(DC))$. The convergence of posterior expectations ensures that $\alpha(CD)$ converges to zero and $\alpha(DC)$ to $F(0)/F(\alpha_1)$. The ranking of the functions, for small $\Theta_C$ and $\Theta_D$, then follows from the observation that $0 < F(0)/F(\alpha_1) < \alpha_1$ (with the second inequality implied by (7)). Uniqueness for the case $\Theta_C = \Theta_D = 0$ is established in Step 3.[31]

*Step* 2 (Applying Proposition 1): The uniqueness of the second-period equilibrium, following mutual first-period cooperation, 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 $\alpha_2$ for which

$$\frac{F(\alpha_2)}{F(\alpha_1)} = \alpha_2. \tag{16}$$

[31] This is the only case (one first-period cooperator and one defector) where uniqueness does not immediately carry over to the case of nonzero $\Theta_C$ and $\Theta_D$. If there are multiple equilibria, they must converge to the unique noiseless equilibrium as $\Theta_C$ and $\Theta_D$ get small, which suffices for our comparative static results.

It suffices for uniqueness that, for any $\alpha_1$,

$$\frac{F(\alpha)}{F(\alpha_1)} = \alpha \quad \Rightarrow \quad \frac{f(\alpha)}{F(\alpha_1)} \leqslant 1. \tag{17}$$

This ensures that the function $F(\alpha)/F(\alpha_1)$ can intersect the diagonal at most once. (The bottom panel of Fig. 10 shows an example where there is no such intersection. Fig. 4 shows a case, requiring $\alpha_1 > 1$, in which there is such an intersection.) Using the first equality in (17) to eliminate $\alpha_1$ from the second, (17) is implied by (9).

*Step* 3 (Establishing uniqueness): To establish uniqueness, it remains to consider continuation play when one agent has cooperated and one has defected in period one. Assumption 1 ensures that there exist $\alpha'$ and $\alpha''$, with

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

with the player $\alpha$ 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 the first-period defector $\alpha$ (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 $\alpha$ will cooperate and those with high values of $\alpha$ will defect (coupled with the observation that agents with dominant second-period strategies will play them).

Now consider the possibilities for the value of $\alpha'$. The penultimate paragraph of the proof of Proposition 2.1 shows, with the help of condition (7), that $\alpha' = 0$ implies that $\alpha'' = \alpha_1$. This combination of $\alpha'$ and $\alpha''$ reproduces the equilibrium of Proposition 2. It thus suffices to show that $\alpha' = 0$ is the only possibility for $\alpha'$.

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

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

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

$$\frac{F(1) - \alpha_1}{1 - \alpha_1} < \alpha_1, \tag{18}$$

or $F(1) < 2\alpha_1 - (\alpha_1)^2$, which follows from (8) and $\alpha_1 > \alpha^*$. Hence, (8) 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 seen that (7) implies $F(0)/F(\alpha_1) < \alpha_1$, given the first of the following inequalities, while (18) gives the second, establishing that (20) cannot hold for $\alpha' = 0$ or $\alpha' = \alpha_1$:

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

If $F$ is linear, then the middle term in (20) is linear in $\alpha'$. 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.

*Step* 4 (Existence): The existence of an equilibrium in the two-period game, for the case in which $\Theta_C$ and $\Theta_D$ are zero, now follows from noting that second-period behavior is both unique and a continuous function of the first-period value of $\alpha_1$. This allows us to apply an intermediate-value argument analogous to that used to establish existence in the one-period game.

Suppose that $\Theta_C$ and $\Theta_D$ are not zero. Notice that utilities in the stage game are bounded (by Assumption 2). Letting $M$ be an upper bound on the absolute value of utility, let $H$ be the set of decreasing functions from $[\underline{\alpha}, \overline{\alpha}]$ into $[0, 2M]$. Then let $(h(\alpha), \rho_1, \rho_{CC}, \rho_{DD}, \rho_{CD}, \rho_{DC})$ denote an element of $H \times [0, 1]^5 \equiv \Upsilon$, interpreted as identifying the value $V(C, \alpha) - V(D, \alpha)$ as a function of $\alpha\,(h(\alpha))$, the probability that a randomly-selected agent cooperates in period 1 $(\rho_1)$, and the probability that an agent who played $x \in \{C, D\}$ in the first period, against an opponent who played $y$, cooperates in the second period $(\rho_{xy})$. Note that $\Upsilon$ is a convex, sequentially compact subset of a metric (and hence locally convex, linear topological) space, with the metric taken to be the sum of the standard Euclidean metric on $[0, 1]^5$ and the metric $||h, h'|| = \int_{\underline{\alpha}}^{\overline{\alpha}} |h(\alpha) - h'\alpha|\, d\alpha$.

We construct the following map. Given $\rho_1$ and the function $h(\alpha)$, we let $\Delta_1(\alpha, \rho_1, h)$ be the unique, increasing function identifying for each value of $\alpha$ the value of the taste shock $\delta$ at which agent $\alpha$ is indifferent between cooperating and defecting in the first period:

$$\Delta_1(\alpha, \rho_1, h) = \pi(D, \rho_1, \alpha) - \pi(C, \rho_1, \alpha) - h(\alpha).$$

We then set first period behavior so that agent $(\delta, \alpha)$ cooperates if and only if $\delta > \Delta_1(\alpha, \rho_1, h)$. This determines a probability $\rho_1'$ with which cooperation occurs in the first period. Next, suppose actions $xy \in \{C, D\}^2$ are taken in the first period. Then we induce second-period behavior by assuming that the agent who chose $x$ assumes that her opponent will cooperate with probability $\rho_{yx}$. The agent in question cooperates in the second period if and only if $\delta > \Delta_{xy}(\alpha, \rho_{yx}) = \pi(D, \rho_{yx}, \alpha) - \pi(C, \rho_{yx}, \alpha)$. In addition, for each period-1 action $x \in \{C, D\}$, the behavior described by $\Delta_1(\alpha, \rho_1, h)$ gives rise to a unique second-period posterior distribution $F_x$ describing beliefs about the type of the agent playing action $x$ in period 1. Together with the induced behavior, these posteriors allow us to calculate probabilities $\rho_{xy}'$ that an agent who played $x$ in the first period, when facing an opponent who played $y$ in the first period, cooperates in the second period. We can also calculate a function $h'(\alpha) = V(C, \alpha) - V(D, \alpha)$.

Noting that the function $h'(\alpha)$ must be decreasing, this defines a mapping from the set $\Upsilon$ into itself. It is immediate that a fixed point of this mapping is an equilibrium of the

two-period game and that the mapping is continuous. The Schauder–Tychonoff theorem (Dunford and Schwartz [12, p. 456]) then implies that the map has a fixed point, and hence the two-period game an equilibrium. □

**Proof of Proposition 3.4** (*Completion*). Suppose that $\alpha_1$ reaches 1, so that the bottom panel of Fig. 5 is relevant. It follows from (16) that $\alpha_2 = \alpha_1$ when $\alpha_1 = 1$, with $\alpha_2$ decreasing as $\alpha_1$ increases above 1. We then have the following list of possible transitions:

$$\begin{array}{llll}
(CD, CD) \to (CC, CD) & (DD, DD) \to (CD, DD) \\
(CD, DD) \to (CC, CD) & (DD, DD) \to (DC, DD) \\
(CD, DD) \to (CC, DD) & (CC, CC) \to (CC, DD) \\
(DC, DC) \to (CC, DC) & (CC, CC) \to (CC, CD) \\
(DC, DD) \to (CC, DC) & (CC, CC) \to (CC, DC) \\
(DC, DD) \to (CC, DD) & (CC, CD) \to (CC, DD) \\
(DD, DD) \to (CC, DD) & (CC, DC) \to (CC, DD).
\end{array} \tag{22}$$

Again, the conclusion is that the incidence of $(CD, CD)$, $(DC, DC)$ and $(DD, DD)$ falls. Hence, each of these three outcomes declines as $\lambda$ increases. However, we now also find that $(CC, CC)$ decreases, leading to the observation that $(CC, CC)$ increases for $\lambda$ such that $\alpha_1(0, \lambda) < 1$ and decreases for $\lambda$ such that $\alpha_1(0, \lambda) > 1$. Next, (22) shows that $(CC, DD)$ increases in $\lambda$ 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 $\lambda$ with $\alpha_1(0, \lambda) > 1$, and, from (22), that there are transitions both to and from these outcomes as $\lambda$ increases when $\alpha_1(0, \lambda) > 1$, leading to an ambiguous comparative static. □

## Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jet.2004.09.002 and http://www.ssc.wisc.edu/~andreoni/WorkingPapers/AndSamInstr.htm.

## References

[1] J. Andreoni, J.H. Miller, Giving according to GARP: an experimental test of the consistency of preferences for altruism, Econometrica 70 (2002) 737–753.
[2] J. Andreoni, J.H. Miller, Rational cooperation in the finitely repeated prisoners' dilemma: experimental evidence, Econ. J. 103 (1993) 570–585.
[3] J. Berg, J. Dickhaut, K. McCabe, Trust, reciprocity, and social history, Games Econ. Behav. 10 (1995) 122–142.
[4] K. Binmore, C. Proulx, L. Samuelson, J. Swierzbinski, Hard bargains and lost opportunities, Econ. J. 108 (1998) 1279–1298.
[5] M. Blonski, D.A. Probst, The emergence of trust, Economic working paper, University of Mannheim, 2001.
[6] G.E. Bolton, A. Ockenfels, ERC: a theory of equity, reciprocity and competition, Amer. Econ. Rev. 90 (2000) 166–193.
[7] C. Camerer, Behavioral Game Theory: Experiments in Strategic Interaction, Russell Sage Foundation and Princeton University Press, Princeton, 2003.

[8] C. Camerer, T.-H. Ho, J.-K. Chong, Sophisticated experience-weighted attraction learning and strategic teaching in repeated games, J. Econ. Theory 104 (2002) 137–188.
[9] C. Camerer, K. Weigelt, Experimental tests of a sequential equilibrium reputation model, Econometrica 56 (1988) 1–36.
[10] S. Datta, Building trust, STICERD Discussion Paper 96/305, London School of Economics, 1996.
[11] D.W. Diamond, Reputation acquisition in debt markets, J. Polit. Econ. 97 (1989) 828–862.
[12] N. Dunford, J.T. Schwartz, Linear Operators Part I: General Theory, Wiley, New York, 1988 (Wiley Classics Library Edition).
[13] A. Falk, U. Fischbacher, A theory of reciprocity, CEPR Discussion Paper 3014, University of Zurich, 2001.
[14] E. Fehr, K.M. Schmidt, A theory of fairness, competition and cooperation, Quart. J. Econom. 114 (1999) 817–868.
[15] D.M. Kreps, P.R. Milgrom, J. Roberts, R.J. Wilson, Rational cooperation in the finitely repeated prisoners' dilemma, J. Econ. Theory 27 (1982) 245–252.
[16] D.M. Kreps, R.J. Wilson, Reputation and imperfect information, J. Econ. Theory 27 (1982) 253–279.
[17] D.K. Levine, Modeling altruism and spitefulness in experiments, Rev. Econ. Dynam. 1 (1998) 593–622.
[18] R.D. McKelvey, T.R. Palfrey, An experimental study of the centipede game, Econometrica 60 (1992) 803–836.
[19] P.R. Milgrom, J. Roberts, Predation, reputation and entry deterrence, J. Econ. Theory 27 (1982) 280–312.
[20] M. Rabin, Incorporating fairness into game theory and economics, Amer. Econ. Rev. 83 (1993) 1281–1302.
[21] A.E. Roth, J.K. Murnighan, Behavior in repeated play of prisoner's dilemma, J. Math. Psychol. 17 (1978) 189–198.
[22] J. Watson, Starting small and renegotiation, J. Econ. Theory 85 (1999) 52–90.
[23] J. Watson, Starting small and commitment, Games Econ. Behav. 38 (2002) 176–199.