

Social Preferences in the Prisoner's Dilemma Game: An Evolutionary Analysis

Anders Poulsen and Odile Poulsen
Department of Economics
The Aarhus School of Business
Prismet
Silkeborgvej 2
DK-8000 Aarhus C
Denmark

December 2, 2002

Abstract

The conventional assumption in economics, that individuals have 'materialistic' preferences, has been questioned by experimental evidence. In this paper we allow players to have various kinds of 'social' preferences and explore, in an evolutionary model, how well these players perform in the one-shot Prisoner's Dilemma. We show that if players have enough information about other players' preferences, then reciprocity always survives, whereas this is not true for materialism. Altruism survives too, and this is actually harmful for cooperation. We analyze both the simultaneous, sequential and a 'mixed' Prisoner's Dilemma game.

Keywords: One-shot Prisoner's Dilemma game; simultaneous/sequential moves; social preferences; indirect evolutionary approach; reciprocity; altruism; materialism; evolutionary stability.

JEL Classification: B41; C72; D74; Z13.

1 Introduction

The conventional assumption in economics is that individuals have 'materialistic' preferences, i.e., their behavior is solely motivated by their material self-interest. This 'materialism' hypothesis has, however, been questioned by experimental evidence. Here players often behave in a cooperative, or fair, way that does not seem compatible with a self-interested preference. There is, for example, often co-operation in a (sequence of) one-shot Prisoner's Dilemma (PD) games (see e.g. Dawes and Thaler (1988)).

The same is true for situations where players must contribute to a public good (see Ledyard (1995)). In bargaining experiments with games like the 'ultimatum game' (see Güth et. al. (1982)) fairness and reciprocity norms seem to affect the outcomes (see Roth (1995)). Moreover, there are enough examples from everyday life: People return lost wallets; they tip the waiter even when being sure to never visit the same restaurant again; they give to charity, and so on.

Indeed, experimental research has documented the existence of a significant fraction of players with 'social preferences'. Following Fehr and Fischbacher (2002), we say that a player has social preferences when she cares not just about her own money payoffs, but also cares about the entire distribution of money payoffs between her and other reference agents, and cares about *how* this distribution was brought about. We refer the reader to e.g. Fehr and Schmidt (2001) for a survey of the experimental findings. There are several important kinds of social preferences: Reciprocity, inequity aversion, altruism and spiteful preferences.

Reciprocity may be characterized as being 'kind' toward someone who is, or is expected to be, kind and to be 'unkind' toward someone who is (expected to be) unkind. Reciprocity is, in other words, *conditional* niceness.¹ What is important in this paper is the fact that when two reciprocally motivated people meet, there are several possible outcomes: One is where both individuals are kind, and the other is where both are unkind. We will, unlike most of the existing literature (see below), allow for the possibility that reciprocal players, because of mistakes or misunderstandings, sometimes end up not both being kind.

Whereas reciprocity is concerned with the decision process, inequity aversion (see Fehr and Schmidt (1999) and Bolton and Ockenfels (2000)) is only concerned with the outcomes themselves. An inequity averse person prefers an equitable distribution of resources. Such a person is willing to increase the opponent's material payoff if the current distribution of resources is too favorable for the person. Conversely, if the opponent is about to get too much, the inequity averse person is willing to take actions that decrease the opponent's material payoff. As we will see, reciprocity and inequity aversion often work in the same direction, since an action of the opponent that will establish an inequitable distribution of resources is naturally interpreted by a player as being 'unkind'.

Unlike the reciprocal person, an altruist is *unconditionally* nice. The key difference between a reciprocally and an altruistically motivated person is that the latter will never take actions that decreases the opponent's payoff. That is, an altruist will never 'punish' the opponent. Finally, we have spiteful, or envious, preferences. A player with these preferences values the opponent's payoff negatively and seeks to maximize the difference between his own and the opponent's payoff - even when this means giving up some of his own material payoff.

We would like to explain why some people have acquired these social preferences, and why other people have purely materialistic preferences. In this paper we therefore endogenize the preferences individuals have. Will they be materialistic, or social - and, if the latter is the case, what kinds of social preference? To do this, we let players with different preferences 'compete' against each other, in order to determine

¹For good surveys, see Fehr and Gächter (2000) and Sethi and Somanathan (forthcoming).

what preference(s) will emerge as the 'winner' of the economic 'struggle for survival'. This methodology is called the 'indirect evolutionary approach' (see e.g. Güth and Yaari (1992) and Güth (1995)): Players act rationally given their preferences, but those preferences may change over time, as the result of a socioeconomic or cultural learning and imitation process. The key assumption is that preferences who give their 'users' higher-than-average *money* payoffs tend to be adopted by more players over time. An interpretation is that, in order to survive in the economic system, one needs to perform *materially* well. However, this assumption, that 'only money matters', does not a priori bias the analysis towards the survival of materialistic preferences: *Any* sort of preference that leads players to a materially superior, or just reasonable, behavior will prosper. Thus, if players with, say, reciprocal, preferences earn more money than those with materialistic preferences, the population frequency of the former will tend to increase at the expense of the latter.

We wish to examine the survival ability of social preferences in a quite harsh setting. To do this, we use the one-shot Prisoner's Dilemma as the basic game. The environment in which this game is played is varied along two dimensions: The game form (simultaneous-move, sequential or 'mixed') and information (perfect information and no information about other players' preferences). The main contribution and result of the paper consists, we believe, in showing the following: The more information players have about opponent's preferences, the more likely it is that reciprocity and altruism, and hence cooperation, will survive.

More precisely, under perfect information reciprocal, materialistic and altruistic preferences co-exist in the simultaneous-move PD game, with their population proportions permanently fluctuating. Such a co-existence result has not, to our knowledge, been seen before. In the sequential PD game, the only stable preferences are those that are reciprocal and/or altruistic. There are no players with materialistic preferences. We also study a 'mixed' game, where players are sometimes engaged in the simultaneous, and at other times, in the sequential game. Here evolutionary preference selection leads to a unique asymptotically stable population, where all three preference types co-exist. Fluctuations in population proportions eventually die out.

These results should, in our opinion, be interpreted as an indication, in an evolutionary context, that the standard assumption that players *always and only* have materialistic motivations, is ill-founded and inappropriate. Such a statement would need to be qualified: Certain environments are more conducive for social preferences than others. The key features that we have stressed here are the amount of information that is available about other people's motivations and the exact way people interact (the move protocol).

The requirement, mentioned above, that players must have enough information about other players' preferences, is crucial: It allows two reciprocal players to perform well enough against another. And, equally importantly, it allows a Reciprocator to avoid being 'exploited' by Materialists. If, on the other hand, players can not recognize their opponent's motivations, then reciprocity and altruism can not thrive in the population. Our analysis shows that in communities with anonymous interaction, where players know little about each other, reciprocity and altruistic behavior should not be expected to emerge. If, on the other hand, it is possible for players to acquire information about opponents prior to interaction, then reciprocity and/or altruism is

a real possibility, and sometimes the only possibility. This also shows, in our opinion, that the standard assumptions of materialistic preferences *and* perfect information are inconsistent.

But, how is it that a player can know whether another person is, say, a reciprocal or materialist type? Strictly speaking, of course, this is impossible: Individuals cannot 'see' inside other persons' heads. However, several authors, such as Robert Frank (Frank (1988)) have argued that it is often possible to correctly deduce people's characteristics, and underlying motivations, from physical tell-tale signs, such as facial expressions and body posture. Another possibility is that individuals have access to information about an opponent's previous behavior, from encounters with other people. Given this information, people form an opinion about what can be expected from the opponent.² Yet another possibility is that individuals base their evaluations of other individuals' preferences using indicators such as income, skin color, area of residence, and so on. We leave for future research the task to incorporate such, and other, realistic features into models of preference evolution.

Do our results then *explain* the *causes* for the many instances of co-operation in experiments and in everyday life in Prisoner's Dilemma like situations? We believe the answer is a 'yes', but in an indirect way. First, in most experiments players do not have information about fellow players' preferences. Interaction is deliberately kept anonymous. It follows that our results, emphasizing the role of information about preferences for reciprocity and altruism to survive, do not directly apply to the experimental findings. What happens in the lab instead, it is probably reasonable to say, is that subjects believe that there are sufficiently many reciprocal and altruist people out there. Given these (often correct) beliefs, and their reciprocal/altruistic preferences, these subjects optimally cooperate in the experimental lab.

The question, therefore, is: How is it that some (often a significant fraction of) subjects have these beliefs and preferences? It is this question that our model supplies some answers to: The beliefs and preferences have been shaped in the outside 'Game of Life' and are consequently used in the experimental lab, too. The Game of Life is such that players with social preferences do survive, side by side with materialistic players. In our model not all players could be materialists in the Game of Life, for reciprocal players could invade. Similarly, not everybody could be reciprocal, or altruist, since other player types would outperform them. The result is thus co-existence between players with different preferences. This is why some players, when seated in the lab in an anonymous setting, while others do not.

There are three key differences between our model and the existing literature.

First, the existing papers typically only allow for two preference types that 'compete' against each other. We allow for up to four preference types. This is important: Any restrictions on the number of preferences that evolution can work with means that the evolutionarily successful preferences in the restricted model may be different from those that would occur if people were allowed to develop more kinds of preferences.³ Indeed, we show that the models with only the Reciprocator and the Materialist types available may give too optimistic predictions about the occurrence of cooperation, due

²See also Harrington (1995) and Kandori (1992).

³Sethi and Somanathan (forthcoming) observe the same for other models of reciprocity.

to their exclusion of players with altruistic preferences.

Second, we take a more detailed look at how reciprocal players perform against each other. As already mentioned, when two Reciprocators meet, the outcome where each player defects is an equilibrium. This possibility is, so to speak, the 'dark side' of reciprocity. It is the conditional niceness/nastyness that distinguishes reciprocity from altruism. Thus, rather than simply assuming that two reciprocal individuals manage to cooperate, which is done in most of the existing literature, we allow for the possibility that they sometimes end up playing defect. This has the crucial implication that players with altruistic preferences survive.

Third, we pay special attention to how the Prisoner's Dilemma game is played: Do the players make choices simultaneously, or does one player make a choice before the other? Or, are the players sometimes engaged in the simultaneous, and other times in the sequential game? We show this makes a difference for the preferences that survive. The reason is that in the sequential move game it is easier for reciprocally minded players to co-ordinate their actions, and, moreover, easier for reciprocal players to 'punish' materialistic players. This 'disciplining' effect of reciprocity on materialistically motivated players has been observed in other contexts, see e.g. Fehr and Schmidt (1999).

We now turn to a more precise description of the literature dealing with preference evolution in the Prisoner's Dilemma. Ockenfels (1993) shows that cooperation is evolutionarily stable as long as players are sufficiently likely to recognize other players' preferences; this recognition allows the Reciprocators to 'reward' other Reciprocators and to 'punish' the Materialists. However, Ockenfels only studies a subset of the preferences types we study. Furthermore, he simply assumes that two reciprocators always manage to co-ordinate on the equilibrium where each player co-operates.

In Guttman (2000) two preference types compete against each other: A Reciprocator type and a Materialist type; there is no Altruistic type in his analysis. Guttman shows that the Reciprocator type drives the other type to extinction, as long as Reciprocators can recognize the Materialists. However, his model is open to the same criticism as Ockenfels' model: There are few admissible preference types and it is simply assumed that two Reciprocators always manage to co-ordinate on the equilibrium with co-operation.

Fershtman and Weiss (1998) explore the role of social status in bringing about cooperation in the PD game. In their model, individuals get status from being more cooperative than the average and individuals care to a differing extent about status relative to money payoffs. They show that for cooperation to establish itself, the concern for social status must be sufficiently strong (such that socially minded people can co-ordinate on the cooperative outcome), but not too strong (then socially minded individuals act like Altruists who can be exploited by asocial people). However, it is, once more, relevant to point out that Fershtman and Weiss only allow for two preference types. Indeed, in their model the set of feasible types is either the Materialist and the Reciprocator type, or the Materialist and the Altruist type.

Whereas our model, and the papers cited above, deal with the one-shot Prisoner's Dilemma, preference evolution has also been analyzed in the (finitely) repeated Prisoner's Dilemma game. Guttman (1999) study interaction between two types, an

'opportunist' (the standard player type) and a 'reciprocator'. Among the features of the model is that players choose their partners and that the past history of players is known. Guttman shows that if the number of periods is sufficiently large, and players observe a signal that has some correlation with the opponent's type, then both types co-exist in the evolutionarily stable outcome. The reciprocators always cooperate, while the opportunists cooperate up to, but not including, the last period.

There are other models of preference evolution. We refer the reader to Ely and Yilankaya (2001), Ok and Vega-Redondo (2001) and Sethi and Somanathan (2001). Our result, that there must be enough information about fellow players' preferences for social preferences to emerge has also been established, in somewhat different contexts, in Güth (1995) and Ok and Vega-Redondo (2001).

The result that different player types can co-exist in an evolutionarily stable outcome has also been observed in (direct) evolutionary models, where players are 'hard-wired' to a certain behavior. See, for example, Amann and Yang (1998), Sethi (1996) and Vogt (2000). This evolutionary approach is complementary to, but conceptually very different from the indirect approach used here.

2 The Model

2.1 The Prisoner's Dilemma Game

Our PD game has the following *money* payoffs:

	<i>C</i>	<i>D</i>
<i>C</i>	1	<i>b</i>
<i>D</i>	<i>a</i>	0

where $a > 1 > 0 > b$ and '*C*' and '*D*' stand for 'Co-operate' and 'Defect', respectively. If both players only care about their own money payoffs, the unique outcome is (D, D) , which in terms of money is worse than (C, C) .

2.2 Preferences

A player is characterized by a strict preference ordering over the four outcomes. Let (i, j) , where $i, j = C, D$, denote the outcome where a player chooses i and the opponent chooses j . However, we are only interested in the behavior that a preference ordering generates. Thus we care about the (pure) *best replies*, derived from the ordering. That is, what will a player choose, given that the opponent plays C , and what will he choose, given the opponent plays D ?

The **Materialist** (M) preference type: D is strictly dominant. The **Reciprocator** (R) preference type: Play C if the opponent plays C and play D if the opponent plays D . That is, the outcome (C, C) is strictly preferred to (D, C) and (D, D) is strictly preferred to (C, D) . The **Altruist** (A) preference type: C is strictly dominant.

In Section 6 below, we consider the fourth preference type.

2.2.1 Interpretation

The 'Materialist', 'Reciprocator' and 'Altruist' labels reflect the following interpretation of what kinds of outcomes of the PD game they would like, and why: The Materialist seeks to maximize his monetary return; the Reciprocator perceives a choice by the opponent to defect as 'unkind' and hence chooses to defect, too; an act of cooperation by the opponent is perceived as 'kind' and hence the person cooperates, too. The Altruist is determined to try to establish a cooperative outcome and will always do her part, independently of what the opponent does. Alternatively, she prefers to act to maximize the opponent's return.⁴

However, it should be stressed that other interpretations are possible. First, we could just as well have used the label 'Inequity averse' instead of 'Reciprocator'. This is because a player with inequity averse preferences would do the same as the Reciprocator: If the opponent plays C , he plays C as well, i.e., he prefers to avoid getting the lion's share; and if the opponent plays D , the player chooses D , too, in order to avoid being a sucker. Similarly, instead of 'Materialist' we could use 'Spiteful' (or 'Envious'); this is because in the Prisoner's Dilemma game maximizing one's own money return is the same as minimizing both the opponent's relative and absolute money payoff. Both requires defection no matter what the opponent does. Under this interpretation the 'Materialist' type actively compares his payoffs with other players, and is just as 'social' as the other preference types.

Lastly, the following 'minimalist' interpretation of preferences is perfectly possible: Players just have a preference ordering and do not know 'why' they prefer what they prefer (that is, if we asked them why they have the preferences they do, they would not know what to answer). This is perhaps actually the interpretation that is most in line with evolutionary thinking. Our labels 'Materialist' etc., are then completely arbitrary are for convenience only.

2.2.2 Other (but Potentially more Restrictive) Ways to Model Social Preferences

An alternative to our approach is to write down a utility function, defined over the monetary consequences. From this utility function one may then derive the best replies we have directly postulated above. However, in this section we argue that the utility function approach is more restrictive than our approach, and may be so in important ways. The reason is that if one constrains oneself from the start by working with a specific class of utility functions, then the best replies one get may be only a subset of the ones we have directly postulated. One then works with only a subset of the possible preference types.

A popular utility function that has been used to model reciprocity, is the form suggested in Fehr and Schmidt (1999). Suppose an outcome is realized that gives

⁴Or, possibly, the sum of the players' monetary return. This requires that the parameters a and b satisfy $0 < a + b < 2$. Our results below do not depend on whether this is the case or not.

money payoffs (π_1, π_2) to players 1 and 2. Then player i , where $i = 1, 2$, gets utility

$$u_i(\pi_i, \pi_j) = \pi_i - \alpha_i \max\{\pi_j - \pi_i, 0\} - \beta_i \max\{\pi_i - \pi_j, 0\},$$

where $\alpha_i > 0$ and $\beta_i > 0$ represent player i 's disutility from disadvantageous and advantageous inequality, respectively. These preferences are said to represent 'inequality aversion'.

Depending on the parameter values α_i and β_i different best replies for our PD-game emerge. If the opponent plays C , an individual plays C too, when β_i is sufficiently large. However, if the opponent plays D , a player *always* plays D .

Thus we can only get the best reply behavior of the Reciprocator and the Materialist type, but never that of the Altruist type.

Another model is Bolton and Ockenfels (2000). Let σ_i denote player i 's share of the total monetary payoff: $\sigma_i = \pi_i/(\pi_1 + \pi_2)$ if $\pi_1 + \pi_2 > 0$ and $\sigma_i = 1/2$ if $\pi_1 = \pi_2 = 0$, where $i = 1, 2$.

In the simplest formulation of their model a player has preferences

$$u_i(\pi_i, \pi_j) = \gamma_i \pi_i - (1/2) \delta_i [\sigma_i - 1/2]^2,$$

where $\gamma_i \geq 0$ and $\delta_i \geq 0$. Thus subjective payoff depends positively on the player's monetary payoff but is diminished whenever an unequal distribution of money arises. Bolton and Ockenfels assume that all monetary payoffs are positive. We therefore, without any loss of generality, add $-b$ to all the payoffs in our PD game. We may then verify that if the opponent plays C , a player chooses C when δ_i is sufficiently large relative to γ . However, a player *always* plays D if the opponent plays D . As in Fehr and Schmidt's specification, we can only generate the behavior corresponding to the Materialist and the Reciprocal type.

We believe these examples show that our approach, of writing down a set of best reply patterns (or, alternatively, of postulating a set of preference orderings generating some best reply patterns), is never more restrictive, and (as just shown) is typically more general, than postulating a specific class of utility functions. Doing the latter may exclude some behavior that may actually turn out to be quite viable in the evolutionary model.

2.3 Evolutionary Selection

We assume there is a large population of players and that at each instant of time the players are randomly matched in pairs. Each pair of players then play the PD once. Players are then re-matched and the process is repeated indefinitely.

Let x_i , with $i = A, R, M$, denote the population fraction of players of type i , where $0 \leq x_i \leq 1$ and $\sum_i x_i = 1$. Then $x = (x_A, x_R, x_M)$ is the population state. Denote by $\pi(i, x)$ the expected *money* payoff to a type i player and let $\pi(x, x)$ denote the average

expected payoff at the population state x . Then the evolution of the population share of players of type i is given by

$$\dot{x}_i = x_i[\pi(i, x) - \pi(x, x)].$$

This is the well-known Replicator Dynamic (Taylor and Jonker (1978)). It says that the growth rate of players with preference $i = A, R, M$ grows if these players earn above-average *money* payoff. We wish to describe the dynamic of preference evolution and to find those population states that are (asymptotically) stable for this dynamic.

3 Simultaneous Interaction

We first assume that two matched individuals play a simultaneous-move PD game, i.e., each player chooses C or D without knowing the opponent's choice. We assume that the players' preferences are common knowledge. This assumption is standard in game theory and we retain it until Section 3.3 below.

For the evolutionary analysis we need to compute the money payoffs π_{ij} , where $i, j = A, R, M$, obtained by a player with preference i when she is matched with an opponent of type j . These payoffs are given in the matrix below:

	A	R	M
A	1	1	b
R	1	π_{RR}	0
M	a	0	0

Figure 1: The money payoffs in the evolutionary game with three preference types; A = Altruist; R = Reciprocator; M = Materialist.

Consider, for example, a meeting between an M -type and an R -type. The M -type always plays D and the R -type consequently plays D , too. Thus the money payoff to each player is the mutual defection payoff, zero: $\pi_{MR} = \pi_{RM} = 0$. Similarly, in an encounter between an A -type and an R -type, the former always plays C and the R -type then responds with C , too. Thus $\pi_{AR} = \pi_{RA} = 1$.

In a meeting between two R -types, there are two possible outcomes, corresponding to the two strict Nash equilibria: (D, D) and (C, C) .⁵ If the players could co-ordinate on the (C, C) Nash equilibrium their money payoff would equal 1, while playing the (D, D) Nash equilibrium would give each player zero. We make the following assumption:

Assumption 1 *The money payoff that an R -type earns when meeting another R -type, π_{RR} , satisfies*

$$0 < \pi_{RR} < 1.$$

⁵There is also a symmetric and mixed Nash equilibrium - see Section 3.2 below.

Assumption 1 implies that a Materialist performs worse than a Reciprocator against another Reciprocator. It also implies that two Reciprocators perform better than two Materialists, but not as well as two Altruists. In Section 3.2 below, we take a closer look at when this assumption is likely to hold, and what happens when it does not.

Looking at the payoff matrix above reveals that an all-Altruist population is invaded by Materialists: The Altruist incumbents always cooperate, also when they meet a Materialist mutant, so the mutant performs better than the incumbents and consequently spreads in the population. Similarly, an all-Materialist population can be invaded by Reciprocators: In an all- M population an R mutant gets zero, like the incumbents, but the mutant performs better than an incumbent against another R mutant. Thus the R mutants spread in the population. Finally, an all-Reciprocator population is invaded by *altruistic* mutants. This happens because whereas the reciprocal incumbents sometimes end up defecting in each other, an altruist is always met by cooperation.

What kind of preferences are then represented in a stable population? The following proposition gives the answer:

Proposition 1 *Consider the evolutionary game with money payoff matrix given in Figure 1 above. Then:*

- (a). *Any population where all players have the same preference is unstable.*
- (b). *Suppose all three preferences are represented in the initial population. Then any future population will also contain players with all three preferences.*
- (c). *More precisely, there is a unique interior equilibrium population, x^* , given below, where all three preferences are present. This equilibrium is a center, i.e., it is surrounded by periodic orbits. Thus, whenever the population is not initially exactly at x^* , the population frequencies of the preference types fluctuate endlessly.*

$$x^* = [x_A^*, x_R^*, x_M^*] = \left[\frac{-b\pi_{RR}}{D}, \frac{b(1-a)}{D}, \frac{(1-\pi_{RR})(a-1)}{D} \right], \quad (1)$$

where $D = (1-b)(a-1+\pi_{RR}) - a\pi_{RR}$.

Proof: The proof follows by an application of the results in Bomze (1983). We refer the reader to the Appendix.

Why do we get these never-ending fluctuations? First of all, there is what one might call a 'cyclic relationship' between the three preference types: Against preference R , preference A performs best; and against preference A preference M is optimal; and facing preference M , both preference M and R are optimal, but preference R performs better against any mix of the two; this leads to preference A being optimal again, and so on. This leads to the never-ending fluctuations in the population shares. As the share of preference R increases, the share of A -types increases, while the share of M -types fall; however, once there are sufficiently many A -types, preference M can gain foothold; this lowers the share of A and R -types, after which type R again gains territory, and so on.

A natural question here is: Why do the fluctuations not dampen, or explode? The reason is the following feature of interaction: Facing a Materialist type a Reciprocator and a Materialist type perform *exactly* as well (they both defect and so get zero payoff). A mathematician would say that this payoff tie, $\pi_{RM} = \pi_{MM}$, in the payoff matrix in Figure 1 is a 'non-robust' feature of the model, in the sense that a small perturbation in these payoffs will change the dynamic (see also Zeeman (1980)). One answer to such objections is that our payoffs in the evolutionary game are *endogenous*: They are derived from the optimal behavior of the players, and so it does not really make sense to discuss arbitrarily small changes in these payoffs.⁶ Instead one should discuss the robustness of the preference dynamic to underlying changes in the 'deeper' parameters of the problem. Below we explore the role of 'perturbing' two such environmental parameters, namely the amount of information players have about each other and the move structure itself: Do players move simultaneously or sequentially?

The result that a population of reciprocally minded players is unstable, since an Altruist can invade, has not, to our knowledge, been observed in other models of preference evolution.⁷ Once more, it follows because we do not simply assume that reciprocators can always establish full cooperation. The result highlights some of the evolutionary costs and benefits that are associated with having different preferences: The benefit of having reciprocal preferences is that one can prevent other players (especially Materialists) from taking advantage of oneself: A Reciprocator will always defect when meeting a Materialist. The cost of reciprocity is that there is a risk that two reciprocators both play defect. However, there is another problem: A Reciprocator always cooperates with an Altruist and this is actually harmful for the population as a whole. To see this, suppose the reciprocal players defected against the Altruists. Then it is not difficult to see that the population consisting only of reciprocal types would be evolutionarily stable.⁸ Then we would only get beneficial *C*-choices, and not the *C* and *D* choices in the fluctuating populations given in Proposition 1 above. In other words, any ability of the Reciprocators to be 'nasty' against *both* Altruists and Materialists would actually be beneficial, in the sense of ensuring that only a population of reciprocal players would be evolutionarily stable. The 'problem', however, for reciprocity is that, at least in our simple model, it cannot do both.

The cost of being an Altruist is that one is easy prey for Materialistic adversaries. However, the benefit is that one is sure to always establish cooperation when meeting other Altruists and Reciprocators. In particular, the altruist performs better against Reciprocator players than a Reciprocator player. Thus, even though the Altruist preference type may seem quite naive and defenseless, it nevertheless survives with the other preference types. And, as just seen above, the presence of altruism is actually counterproductive for establishing co-operation.

⁶In Robson (1992) a similar argument is presented.

⁷But see Binmore and Samuelson (1992) for a similar kind of result, although in a different context.

⁸On deleting the 'A' row and 'A' column in the matrix for the evolutionary game (Figure 1), the *R* strategy is a strict Nash equilibrium and hence evolutionarily stable. See Weibull (1995).

3.1 Comparison with the Classical 'Direct' Evolutionary Analysis

In the classical evolutionary analysis of the repeated PD game, players do not maximize utility but are behaviorally 'programmed' to certain behaviors (see e.g. Axelrod (1984) and the survey in Sethi and Somanathan (forthcoming)). Among the strategies that have attracted a lot of attention are the following three: 'Always Defect', 'Always Cooperate' and 'Tit-for-Tat', namely start out cooperating and subsequently do what the opponent did at his previous move. The Tit-for-Tat and Always Cooperate strategies can co-exist side by side, since they all cooperate with each other. Such populations are stable as long as an Always Defect player cannot invade. This requires that there are not too many Always Cooperate players in the population. If, however, a certain critical proportion of Always Cooperate players is exceeded, players who always defect invades⁹ and the population is taken towards the population consisting of only Always Defect players. This population is stable.

Our Materialist, Reciprocator and Altruist preference type look quite similar to the 'Always Defect', 'Tit-For-Tat' (TFT) and 'Always Cooperate' strategies. Indeed, a player who is 'programmed' to always defect (or co-operate) is behaviorally indistinguishable from a player with preferences that make defection (co-operation) strictly dominant.

The Tit-For-Tat strategy, however, is behaviorally distinguishable from a Reciprocator player: Two Tit-for-tat players always establish cooperation, but this is not true for two Reciprocators, since they face an equilibrium selection problem and sometimes do not manage to co-ordinate on cooperation. In our context an Altruist performs better than a Reciprocator in a population of Reciprocators, whereas an Always Cooperate player performs exactly as well as a Tit-For-Tat player in a population of Tit-For-Tat players. It is this difference between reciprocity and the Tit-For-Tat strategy that makes our indirect analysis different from the classical 'direct' evolutionary analysis.

3.2 A Closer Look at the Interaction between two Reciprocators

Above we assumed that two Reciprocators earned a monetary payoff that was between the pure defection and co-operation payoffs (cf. Assumption 1). How plausible is this assumption? In this section we seek to endogenise the monetary payoff that two Reciprocators get. We also analyze what happens when the inequality $0 < \pi_{RR} < 1$ is not satisfied.

As noted earlier, when two R -types meet, they face an equilibrium selection problem: Both (C, C) and (D, D) are strict Nash equilibria. Furthermore, there is also a symmetric mixed Nash equilibrium. We will assume that each R -player plays C with probability λ , where $0 < \lambda < 1$. There are several possible interpretations of this: The first is that the players independently make occasional mistakes in co-ordinating on a Nash equilibrium. Thus λ being close to unity reflects a situation where the R -types

⁹See Sethi and Somanathan (forthcoming) for other examples of this quite general phenomenon.

are almost able to co-ordinate on the (C, C) Nash equilibrium; and the situation where the players co-ordinate almost always on the (D, D) equilibrium is represented by a very small value of λ .¹⁰ Another interpretation is that two Reciprocators play the symmetric mixed Nash equilibrium and that λ is the probability of playing C .

We therefore have

$$\pi_{RR} = \lambda^2 + (1 - \lambda)\lambda a + (1 - \lambda)\lambda b, \quad (2)$$

and obtain the following characterization:

Proposition 2 *Consider again the evolutionary game with payoff matrix in Figure 1 above. Suppose two Reciprocators each play C with probability λ when they meet. Then:*

(a). *If $0 < a + b < 2$, then $0 < \pi_{RR} < 1$ for all λ . All the results from the previous analysis then continues to hold.*

(b). *If $a + b > 2$, then $\pi_{RR} > 1$ for $\lambda \in (\frac{1}{a+b-1}, 1)$. Then the all-R population is the unique asymptotically stable population.*

(c). *If $a + b < 0$, then $\pi_{RR} < 0$ for $\lambda \in (0, \frac{-(a+b)}{1-(a+b)})$. The all-M population is now the unique asymptotically stable population.*

The proof is in the Appendix.

The condition $a + b < 2$ in part (a) is nothing but the condition that is often imposed on the PD game: $1 > (1/2)(a + b)$.¹¹ The condition $a + b > 0$ says that the net payoff from defection relative to being defected on is positive. Thus, whenever $0 < a + b < 2$, all the results from Section 3 go through.

If, however, $a + b > 2$, as in part (b), it is possible to have $\pi > 1$. This happens if the Reciprocators are sufficiently likely to co-ordinate on the (C, C) equilibrium. They then manage sufficiently often to avoid the inferior zero payoffs from mutual defection and since they get $(1/2)(a + b) > 1$ when not co-ordinating on co-operation, the overall payoff exceeds the cooperation payoff. A Reciprocator player then performs better than any other preference type against its own fellow players, so the all-Reciprocator is the *unique asymptotically stable population*.

Part (c) shows that $\pi_{RR} < 0$ can only happen if $a + b < 0$ and if the players are sufficiently likely to co-ordinate on the (D, D) equilibrium. In this case the players rarely get the high payoff from cooperation and when they do not co-ordinate on defection they get on average only $(1/2)(a + b) < 0$. The overall payoff is then below the payoff from mutual defection. In this case only the all-M population is asymptotically stable.

¹⁰In Fershtman and Weiss (1997), it is assumed that two R-types play a correlated equilibrium: Both players play C with probability λ and both play D with probability $1 - \lambda$.

¹¹The usual interpretation is that it must not pay players in monetary terms to alternate between the (D, C) and the (C, D) outcomes, relative to always realizing the (C, C) outcome.

3.3 When Players have no Information about Other Players' Preferences

In the previous analysis subjective payoffs were common knowledge. It was as if players had their preference types "written" on their foreheads. Let us now instead suppose that a player, when having to decide between cooperating and defecting, receives no information about the opponent's preferences; all interaction is completely anonymous. The only thing a player knows is the *aggregate* distribution of the different preference types in the population.

Anonymous interaction means that a player's preferences can no longer affect an opponent's choice: Two Reciprocators can no longer recognize each other, so they cannot seek to co-ordinate on cooperation. Similarly, when a Reciprocator meets a Materialist the Reciprocator can no longer 'punish' the Materialist by playing defect. Instead all players face the same distribution of preference types and the same distribution of C and D choices. But that implies that the unambiguously best thing to do in terms of money is to defect.

Proposition 3 *Consider the evolutionary game when players do not know their opponents' preferences. Then: In any stable population there are no Altruists and all players defect.*

The proof is in the Appendix.

This proposition underlines the importance of personal communication for reciprocity, and Altruism, to be successful. Players must be able to credibly signal, or communicate, what value system they have, in order to establish cooperative outcomes and in order to punish 'bad' players. In a *completely* anonymous world, cooperation is not possible. Note that the proposition does not say that all players are materialistic in any stable population; in fact, some may be reciprocal. However, in any stable population there are so many Materialists that the Reciprocators defect, too. Thus in any stable outcome materialistic and reciprocal players are indistinguishable from each other. It is the lack of a means of communication that prevents the Reciprocators from 'breaking out' and establishing the cooperation between themselves that would give them an evolutionary advantage over the materialists. In the terminology of Robson (1990), in a completely anonymous world, reciprocal players cannot give each other a 'secret handshake'.¹²

4 The Sequential Prisoner's Dilemma Game

The analysis in the previous sections was based on the assumption that players *simultaneously* made a choice of whether to defect or cooperate. In this section we instead study a *sequential* Prisoner's Dilemma game: When two players are matched, one player is randomly chosen to be first-mover. This player then chooses between

¹²Similar results, in other contexts, about what happens under completely anonymous interaction is given in Güth (1995) and Ok and Vega-Redondo (2001)

cooperate and defect. The other player, the second-mover, observes the first-mover's choice and makes a choice himself. The fact that players are randomly allocated to be first-mover or second-mover is meant to reflect a set-up where, when two players meet each other, random factors decide who moves first. We assume, as before, that players' preferences are common knowledge. In particular, the first-mover knows the second-mover's preferences before the first-mover makes a choice.¹³

Consider a player who as first-mover faces a reciprocal second-mover. If the first-mover chooses cooperate, the second-mover cooperates, too. And if the first-mover were to choose to defect, the second mover would defect as well. Thus the key issue is how the first-mover ranks the (cooperate, cooperate) outcome relative to the (defect, defect) outcome. We will make the following assumption:

Assumption 2 *All players, irrespective of their preference, prefer the (C, C) outcome to the (D, D) outcome.*

We feel this is a plausible assumption, which is entirely in line with our proposed interpretation of materialism, reciprocity and altruism.

We obtain the following matrix for the evolutionary game:

	A	R	M
A	1	1	b
R	1	1	$1/2$
M	a	$1/2$	0

Figure 2: The money payoffs in the evolutionary game for the sequential Prisoner's Dilemma game. A= Altruist; R = Reciprocator; M = Materialist.

The important difference from the simultaneous set-up is that (i). *Sequential interaction allows two Reciprocators to overcome their co-ordination problem* and (ii). *a Reciprocator outperforms a Materialist, both against a Reciprocator and a Materialist opponent.* Observation (i) holds because when a reciprocal first-mover faces a reciprocal second-mover, the first-mover chooses *C* and the second-mover responds with *C*, too (Assumption 2). In a meeting between two Reciprocators, each player therefore earns $(1/2)(1) + (1/2)(1) = 1$. The second observation holds come from the fact that when the Reciprocator is first-mover she effectively *punishes* the Materialist second-mover. And when the Materialist is first-mover she optimally chooses to *cooperate* rather than to defect (Assumption 2 again). The presence of reciprocally motivated players induce materialistic players to behave more cooperatively than they otherwise would. Even though the Materialist prefers to defect, *given* any choice by the opponent, when her own choice will determine the opponent's choice, she is led to cooperate. This is a specific case of a quite general phenomenon: The presence of individuals in the population with reciprocal preferences affects the behavior of materialistically motivated individuals, and, in fact, makes the latter more cooperative. See e.g. Fehr and Schmidt (1999) for a discussion.

The following proposition summarizes our findings for the sequential PD game.

¹³This does not matter for the second-mover.

Proposition 4 *Consider the evolutionary game for the Sequential Prisoner’s Dilemma game, with money payoffs given in Figure 2 above. Then:*

(a). *As in the simultaneous-move game, the all-M population is unstable. However, and unlike the simultaneous-move context, no Materialist players are observed in any stable population.*

(b). *Any population composed exclusively of Reciprocators and Altruists is stable as long as the population share of Altruists, x_A , is sufficiently small: $x_A < 1/(2a - 1)$.*

The proof of this proposition is in the Appendix. In these stable populations all players perform equally well (they play C in all encounters, irrespectively of the preferences of the players). Hence there is no selection pressure. This implies that we have the phenomenon of ‘evolutionary drift’: The population shares of Altruists and Reciprocators may, due to small (and unmodeled) stochastic factors, slowly and unnoticeably change.¹⁴

However, once sufficiently many players have drifted to adopting the Altruistic preference, a Materialist mutant will earn enough to be able to invade the population.

5 A Simultaneous-Sequential Prisoner’s Dilemma Game

In the previous sections interaction in the PD game was either simultaneous or sequential. However, we should not forget that a game model is only intended as a rough approximation to the real interaction. In this section we therefore study the impact of another ‘perturbation’ of the environment: Players are sometimes involved in simultaneous interaction and at other times in sequential interaction. For simplicity, we will assume that, when two players are matched, they engage in the simultaneous game with probability one-half and with remaining probability interaction is sequential. In the latter case this means that a player is equally likely to be first-mover or second-mover. We will refer to this game as the ‘mixed’ PD game. We assume the players know whether they are interacting in a simultaneous or sequential manner.

A player’s evolutionary performance, i.e., his monetary earnings, is now a weighted average of his performance in the simultaneous game and his performance in the sequential one. This gives us the following matrix, where we again assume that $0 < \pi_{RR} < 1$:

	A	R	M
A	1	1	b
R	1	$1/2 + (1/2)\pi_{RR}$	$1/4$
M	a	$1/4$	0

Figure 3: The money payoffs in the ‘mixed’ evolutionary game.

¹⁴We refer the reader to Binmore and Samuelson (1999) for a general analysis of such drift.

Consider a meeting between two R-types. When they play the simultaneous game, a player gets money payoff π_{RR} (cf. Section 3). And when they play the sequential game, he gets money payoff 1 (cf. Section 4). Thus the overall payoff is: $(1/2)\pi_{RR} + (1/2)(1)$. The payoff π_{MR} is computed as follows: With probability one-half the payoff from the simultaneous game emerges, so the R-type earns zero. With probability one-half the sequential game emerges and the R-type gets 1/2. Thus the R-type gets $(1/2)(0) + (1/2)[(1/2)(0) + (1/2)(1)] = 1/4$. Similar reasoning shows that the M-type gets $(1/2)(0) + (1/2)[(1/2)(1) + (1/2)(0)] = 1/4$. Since $\pi_{RR} < 1$, it is still true that $\pi_{AR} > \pi_{RR}$. The new thing is that now *an R-type performs better than an M-type against an M-type*. The implication is given in the following proposition:

Proposition 5 *Consider the evolutionary game where players are equally likely to play the simultaneous PD game and the sequential PD game. Then: There is a unique interior equilibrium population, y^* (given below). This population is asymptotically stable: Every orbit with all three preference types present is attracted to y^* .*

$$y^* = [y_A^*, y_R^*, y_M^*] = \left[\frac{4b + 8\pi_{RR}b - 3}{E}, \frac{4(1 - 4b)(1 - a)}{E}, \frac{8(a - 1)(\pi_{RR} - 1)}{E} \right], \quad (3)$$

where $E = 8\pi_{RR}(a + b - 1) + (4a - 3)(4b - 3)$.

The proof is in the Appendix. An illustration is provided below, again with $a = 2$, $b = -1$ and $\pi_{RR} = 1/2$.

Thus evolutionary selection gives a unique prediction of population behaviour for the 'mixed' game: Any initial population, where all three types are present, will over time evolve to the population y^* .

6 Including the 'Paradoxical' Preference Type

The analysis in Section 3 did not include the following preference type:

The **Paradoxical** (*P*) best reply combination: Play *C* if the opponent plays *D* and play *D* if the opponent plays *C*.

We kept this preference type out for reasons of tractability. However, even though the *P*-preference seems rather odd, we should not a priori exclude it from the analysis. Indeed, since the hallmark of evolutionary analysis is to allow 'traits' to compete against each other and endogenously determine the 'winner', we should not prematurely exclude any competitors from joining the evolutionary 'race'. In this section we therefore consider evolutionary stability when players may also evolve this type of preference.

Rather than first analyzing the simultaneous, the sequential and then the mixed game, we proceed straight to the mixed game, where players are sometimes playing the simultaneous, and at other times play the sequential game (cf. Section 5).

Let, as before, π_{RR} , denote the money payoff to a Reciprocator when meeting another Reciprocator in the simultaneous PD game. Similarly, let π_{RP} denote the money payoff to a Reciprocator when meeting a Paradoxical type in the simultaneous game. The money payoff π_{PR} is defined similarly. The matrix below contains all the payoffs for the evolutionary game.

	A	R	M	P
A	1	1	b	b
R	1	$(1/2)(1 + \pi_{RR})$	$1/4$	$(1/4)[2\pi_{RP} + 1 + a]$
M	a	$1/4$	0	a
P	a	$(1/4)[1 + b + 2\pi_{PR}]$	b	$(1/4)[2\pi_{PP} + a + b]$

Figure 4: The money payoffs in the 'mixed' evolutionary game with four preference types; A = Altruist; R = Reciprocator; M = Materialist; P = Paradoxical type.

Let us explain how we have arrived at the payoffs in the fourth row and column. The first payoff, a , is computed as follows. In the simultaneous game the A type plays C , as usual, and so the P -type plays D . Thus the P -type gets a . In the sequential game the P -type chooses D when first-mover and the A -type responds with C . When the A -type is first-mover, she effectively chooses between (C, D) , giving her money payoff a , and (D, C) , giving money payoff b . We will assume that the A -type is benevolent enough to choose C and so establishes the (D, C) outcome.

When the P -type meets an R -type, there is a unique symmetric and mixed Nash equilibrium in the simultaneous game. In the sequential game a reciprocal first-mover in effect chooses between the outcomes (D, C) and (C, D) . We assume that the R type chooses D and so we get the (D, C) outcome. When the P -type is first-mover, Assumption 2 (Section 4) implies that we get the (C, C) outcome. The P -type therefore gets payoff $(1/2)\pi_{PR} + (1/2)[(1/2)b + 1/2]$. The R -type gets $(1/2)\pi_{RP} + (1/2)[(1/2)a + 1/2]$. When the opponent is an M -type, on the other hand, the outcome is that the P -type chooses C and the M -type chooses D , both in the simultaneous and the sequential game. Finally, suppose two P -types meet. In the simultaneous game there is a unique symmetric mixed Nash equilibrium,¹⁵ giving money payoff π_{PP} . In the sequential game a P -type as first-mover gets a and as second-mover he gets b . Thus the overall monetary payoff to a P -type against another P -type in the mixed game is $(1/2)\pi_{PP} + (1/2)[(1/2)a + (1/2)b]$.

Comparing the performance of the M -type and the P -type shows the following.

Proposition 6 *Suppose $\pi_{PR} < -(1/2)b$. Then:*

- (a). *The M -type weakly dominates the P -type.*
- (b). *When the initial population contains all four preference types, the population proportion of type P players approaches zero as time approaches infinity.*

The proof is in the Appendix.

¹⁵There are also two asymmetric pure Nash equilibria, but we ignore them here.

Part (a) follows from the fact that an M -type always performs strictly better than a P -type against an M and a P opponent; and when $\pi_{PR} < -(1/2)b$, the same is true when the opponent is an R -type. The intuition is the following. When b is sufficiently negative the P -player performs badly against the R -player in the sequential game whenever the P -type is second mover. The P -type's "problem" is the urge to cooperate if the opponent is going to defect and the fact that the P -type, unlike the A -type, does not perform very well against the R -type.

When the proportion of P -type players approaches zero means, as stated in part (b), we can effectively ignore this strategy from the analysis: The population will eventually 'land' on the face of the simplex spanned by the strategies A , R and M , and from then on the dynamic will be as when only these three strategies were available from the beginning. In this case all our results from Section 3, 4 and 5 hold.

We may also remark that were to consider the simultaneous or sequential game in isolation, we would get the following results: In the sequential PD game the P -type *always* disappears with time; this follows because the M -type always weakly dominates the P -type. In the simultaneous game, however, the condition for weak dominance is stricter: It is required that $\pi_{PR} < 0$.

7 Conclusion

In this paper we set up a simple model analyzing what kind of preferences we should expect people to evolve over time when they are engaged in a social dilemma situation of the Prisoner's Dilemma type. This game was played under varying amounts of information about other players' preferences and different move protocols. We showed that when players have information about opponent's preferences, reciprocity will always evolve. Depending on the move protocol, there can be preference heterogeneity, with permanent, or dampening, fluctuations. Our results contribute to providing an evolutionary foundation for the experimentally observed fact that many individuals have social preferences that differ from the materialistic preferences that are normally assumed.

8 Appendix

For simplicity, we set $\pi_{RR} \equiv \pi$.

Proof of Proposition 1:

The equations giving the expected payoffs to the A , R and M preference type are as follows (cf. Figure 1):

$$\pi(A, x) = 1 - x_M + x_M b,$$

$$\pi(R, x) = x_A + x_R \pi,$$

$$\pi(M, x) = x_A a.$$

Solving the system $\pi(A, x) = \pi(R, x)$ and $\pi(R, x) = \pi(M, x)$, using $1 = x_A + x_R + x_M$, gives the solutions from the main text:

$$x_A^* = \frac{-b\pi}{a + b - 1 - ba + \pi - \pi a - b\pi},$$

$$x_R^* = \frac{b(1 - a)}{a + b - 1 - ba + \pi - \pi a - b\pi},$$

$$x_M^* = \frac{(1 - \pi)(a - 1)}{a + b - 1 - ba + \pi - \pi a - b\pi}.$$

We next verify that x_i^* , $i = A, R, M$, are all strictly between zero and one. We note first that all three numerators above are strictly positive. Hence we must verify that the common denominator is strictly positive, too. The denominator can be written as $(1 - b)(a - 1 + \pi) - a\pi$. We observe that, since $a - 1 + \pi > \pi$, this holds whenever $1 - b > a$. Suppose therefore $1 - b < a$. The denominator is strictly positive iff $(1 - b)(a - 1 + \pi) > a\pi$, or $\pi(a + b - 1) < (1 - b)(a - 1)$, i.e.,

$$\pi < \frac{(1 - b)(a - 1)}{a + b - 1}. \quad (4)$$

(since, by assumption, $1 - b < a$, the right hand side is strictly positive). Since always $\pi < 1$, a sufficient condition for (4) is that the right hand side is strictly greater than one. This condition, in turn, is equivalent to $-ab > 0$, which holds since $a > 0$ and $b < 0$. Thus the denominator is strictly positive and we have $x_i^* > 0$ for $i = A, R, M$.

We next verify that the condition $x_A^* < 1$ is equivalent to $\pi < 1 - b$, which always holds. A sufficient condition for $x_R^* < 1$ is

$$\pi(a + b - 1) < a - 1.$$

This clearly holds whenever $a + b - 1 < 0$. Suppose therefore $a + b - 1 > 0$. Then the condition above becomes

$$\pi < \frac{a - 1}{a + b - 1},$$

which is always satisfied, since the right hand side strictly exceeds unity.

We now consider the stability of the interior equilibrium x^* . As already stated in the main text, our proof that x^* is a center builds on the results in Bomze (1983).

Bomze exploits the fact that there is a close relationship between the Lotka-Volterra Dynamic and our Replicator Dynamic (see Hofbauer (1981) and Hofbauer and Sigmund (1998)). Results about the stability of one system may hold for the other system.

First, we may, instead of the matrix in Figure 1, study the equivalent matrix:

	A	R	M
A	0	0	0
R	0	$\pi - 1$	$-b$
M	$a - 1$	-1	$-b$

Or, in abbreviated form,

	A	R	M
A	0	0	0
R	α	β	γ
M	δ	ϵ	θ

Bomze shows (Proposition 6, part (ii)) that iff the quantities $\beta\theta - \gamma\epsilon$, $\alpha\epsilon - \beta\delta$ and $\gamma\delta - \alpha\theta$ all have the same sign, then the Lotka-Volterra dynamic has a unique fixed point, given by $p = \frac{\gamma\delta - \alpha\theta}{\beta\theta - \gamma\epsilon}$ and $q = \frac{\alpha\epsilon - \beta\delta}{\beta\theta - \gamma\epsilon}$. We compute $\beta\theta - \gamma\epsilon = -\pi b > 0$, $\alpha\epsilon - \beta\delta = (1 - \pi)(a - 1) > 0$ and $\gamma\delta - \alpha\theta = -b(a - 1) > 0$.

Hence there is a unique fixed point, (p, q) .

Bomze furthermore shows that if $\beta p + \theta q = 0$, then (p, q) is a center for the Lotka-Volterra system. We have

$$\beta p + \theta q = \frac{b(\pi - 1)(a - 1)}{\pi b} + \frac{-b(1 - \pi)(a - 1)}{-\pi b} = 0.$$

We may therefore conclude that (p, q) is a center for the Lotka-Volterra system. This, in turn, allows us to conclude that our equilibrium x^* is a center for the Replicator Dynamic. ■

Proof of Proposition 2:

Consider the expression

$$\pi_{RR} = \lambda^2 + \lambda(1 - \lambda)a + \lambda(1 - \lambda)b.$$

The condition $\pi_{RR} > 0$ is equivalent to

$$\lambda + (1 - \lambda)a + (1 - \lambda)b > 0 \tag{5}$$

And a sufficient condition for $\pi_{RR} > 0$ is $a + b > 0$, while a necessary condition for $\pi_{RR} < 0$ is $a + b < 0$. Suppose the latter holds. From (5), we that $\pi_{RR} > 0$ when $\lambda > \frac{-(a+b)}{1-(a+b)}$. It then follows that $\pi_{RR} < 0$ when $a + b < 0$ and $\lambda \in \left(0, \frac{-(a+b)}{1-(a+b)}\right)$.

The condition $\pi_{RR} < 1$ is satisfied whenever $a + b < 1$. Suppose therefore $a + b > 1$. $\pi_{RR} < 1$ is equivalent to the expression $\lambda^2[1 - (a + b)] + \lambda(a + b) - 1 < 0$. This holds

whenever $\lambda \leq \frac{1}{a+b}$. Assume this holds. We then verify from (8) that $\pi_{RR} < 1$ whenever $\lambda < \frac{1}{a+b-1}$. This is always satisfied whenever $a+b < 2$. Conversely, we have $\pi_{RR} > 1$ when $\lambda \in (\frac{1}{a+b-1}, 1right)$, for which it is necessary that $a+b > 2$.

Part (a) then follows simply by combining the observations just made above.

Part (b). When $\pi_{RR} > 1$, the R -strategy is a strict Nash equilibrium for the evolutionary game, and hence asymptotically stable for the Replicator Dynamic.

Part (c). To show that the M -strategy is asymptotically stable when $\pi_{RR} < 0$, it suffices to show that it is an ESS (see e.g. Weibull (1995)). Let x^* denote the all- M population. We verify that $\pi(x, x^*) \leq \pi(x^*, x^*)$ for all x . Moreover, $\pi(x, x^*) = \pi(x^*, x^*)$ implies that x uses a mix of strategies R and M . Thus we need to show that for all such $x \neq x^*$ we have $\pi(x^*, x) > \pi(x, x)$. But $\pi(x^*, x) = 0$ and $\pi(x, x) = x_R^2 \pi_{RR} < 0$, so this is satisfied. ■

Proof of Proposition 3:

It is not difficult to see that there can be no Altruists in any stable population. For the Altruists always cooperate, while the Materialists defect, so the latter earns strictly higher expected payoff than the former against any population where some players cooperate. This then implies that in any stable population state all players defect, i.e., they are materialists and/or Reciprocators. ■

Proof of Proposition 4:

Part (a). First, the M -strategy is not a Nash equilibrium and hence unstable. Second, suppose there is a stable population where the M -strategy is present. Then there must be players of type A or of type R , as well. Suppose there are only A and M types in the population. This, however, contradicts stability, since the M -type in such a population earns strictly higher expected payoff than the A -type. A similar contradiction is obtained if only the M and the R type are present. Thus stability implies that all three preference types are present in the population. However, then the R -type earns strictly higher expected payoff than the A -type, again a contradiction of stability.

Part (b). Consider a population, x , with only R and A -players. We then have $\pi(R, x) = \pi(A, x) = \pi(x, x) = 1$. Moreover, we have $\pi(M, x) = x_A a + (1 - x_A)(1/2)$, so $\pi(M, x) < \pi(x, x)$ when $x_A < 1/(2a - 1)$. Then x is a symmetric Nash equilibrium. To show that x is also a Neutrally Stable Strategy (NSS), and hence stable for the Replicator Dynamic¹⁶, we must verify that $\pi(x, x') = \pi(x', x')$ for any $x' \neq x$ using strategy A and R . Since $\pi(x, x') = \pi(x', x') = 1$, x is an NSS (but not an ESS). ■

Proof of Proposition 5:

The equations giving the expected payoffs are now

$$\pi(A, x) = x_A + x_R - x_M b$$

$$\pi(R, x) = x_A + x_R[1/2 + (1/2)\pi] + x_M(1/4)$$

¹⁶We refer the reader to e.g. Weibull (1995).

$$\pi(M, x) = x_A a + x_R(1/4)$$

Solving these equations yields

$$y_A^* = \frac{4b + 8b\pi - 3}{-8\pi + 8b\pi + 9 - 12a - 12b + 16ba + 8\pi a}$$

$$y_R^* = \frac{4(1-a)(1-4b)}{-8\pi + 8b\pi + 9 - 12a - 12b + 16ba + 8\pi a}$$

$$y_M^* = \frac{8(1-a)(1-\pi)}{-8\pi + 8b\pi + 9 - 12a - 12b + 16ba + 8\pi a}.$$

It is not difficult to see that all three numerators in the above expressions are strictly negative. The denominator can be rewritten as $8\pi(a+b-1) + (4a-3)(4b-3)$. Since the last product term is negative, we may conclude that the denominator is strictly negative whenever $a+b-1 < 0$. Suppose then $a+b+1 > 0$. The denominator is then negative iff

$$\pi < \frac{(4a-3)(3-4b)}{8(a+b-1)},$$

A sufficient condition for this to hold is that the right-hand side exceeds unity. This is the same as $(4a-1)(1-4b) > 0$, which holds. Thus the denominator is strictly negative and we have $y_i^* > 0$ for all $i = A, R, M$.

We next verify that $y_A^* < 1$ is the same as $\pi < (3-4b)/2$. This always holds. Similarly, we have $y_R^* < 1$ iff $8\pi(1-a-b) > 5-8a+4b$. The right hand side is always negative, so the condition holds whenever $a+b < 1$. If $a+b > 1$, we have $y_R^* < 1$ iff $\pi < \frac{8a-5-4b}{8(a+b-1)}$. However, this always holds, since the right hand side is strictly larger than one. Finally, we have $y_M^* < 1$ whenever $8b\pi < 4a-1-4b(4a-3)$. This always holds, since the left hand side is negative and the right hand side is positive. Thus we have $0 < y_i^* < 1$ for $i = A, R, M$.

Consider now the dynamic stability of y^* . An equivalent form of the matrix in Figure 3 is

	<i>A</i>	<i>R</i>	<i>M</i>
<i>A</i>	0	0	0
<i>R</i>	0	$(1/2)(\pi - 1)$	$1/4 - b$
<i>M</i>	$a - 1$	$-3/4$	$-b$

where we again use the notation

	<i>A</i>	<i>R</i>	<i>M</i>
<i>A</i>	0	0	0
<i>R</i>	α	β	γ
<i>M</i>	δ	ϵ	θ

We now compute that $\beta\theta - \gamma\epsilon = 3/16 - (1/2)\pi b - (1/4)b > 0$, $\alpha\epsilon - \beta\delta = (1/2)(1 - \pi)(a - 1) > 0$ and

$\gamma\delta - \alpha\theta = (1/4 - b)(a - 1) > 0$. Hence we may again conclude that there is a unique fixed point, (p, q) , where, using the same formula as above, $p = \frac{(1-4b)(a-1)}{3/4-2\pi b}$ and $q = \frac{2(1-\pi)(a-1)}{3/4-2\pi b-b}$. Bomze shows (1983, Corollary 7, part (iiic)) that if $\beta p + \theta q < 0$, then (p, q) , and hence y^* , is asymptotically stable.

We compute

$$\beta p + \theta q = \frac{(1/2)(\pi - 1)(1 - 4b)(a - 1) - 2b(1 - \pi)(a - 1)}{3/4 - 2\pi b - b}.$$

This may be simplified to

$$\beta p + \theta q = \frac{-(1/2)(1 - \pi)(a - 1)}{3/4 - 2\pi b - b},$$

which is strictly negative, since the numerator is strictly negative and the denominator is strictly positive. Hence we may conclude that y^* is asymptotically stable. ■

Proof of Proposition 6:

Part (b): This follows from Weibull (1995), Proposition 3.2: If a pure strategy, i , is weakly dominated by another (mixed) strategy, call it x , then either strategy i approaches extinction over time, or those pure strategies against which x strictly outperforms i , die out. In our case $i = P$ and we may choose x to be the pure strategy M . That is, strategy P is one of those strategies against which x outperforms P . This implies that P dies out over time. ■

9 References

Amann, E. and Yang, C: (1998): "Sophistication and the persistence of cooperation", *Journal of Economic Behavior and Organization*, 37, 91-105.

Axelrod, R.: *The evolution of cooperation*, New York: Basic Books, 1984.

Binmore, K. and Samuelson, L. (1992): "Evolutionary Stability in Repeated Games Played by Finite Automata", *Journal of Economic Theory*, 57, 278-305.

Binmore, K. and Samuelson, L. (1999): "Evolutionary Drift and Equilibrium Selection", *Review of Economic Studies*, 66, 363-393.

Bolton, G. and Ockenfels, A. (2000): "A theory of equity, reciprocity and competition", *American Economic Review*, 100, 166-193.

Bomze, I. (1983): "Lotka-Volterra Equation and Replicator Dynamics: A Two-Dimensional Classification", *Biological Cybernetics*, 48, 201-211.

Dawes, R. and Thaler, R. (1988): "Cooperation", *Journal of Economic Perspectives*, 2, 187-197.

Ely, J. and Yilankaya, O. (2001): "Nash Equilibrium and the Evolution of Preferences", *Journal of Economic Theory*, 97, 255-272.

Gibbons, R. (1992): *A Primer in Game Theory*, Harvester Wheatsheaf, 1992.

Fehr, E. and Falk, A. (2002): "Psychological foundations of incentives", *European Economic Review*, 46, 687-724.

Fehr, E. and Fischbacher, U. (2002): "Why Social Preferences Matter - The Impact of Nonselfish Motives on Competition, Cooperation, and Incentives", *Economic Journal*, 112, C1-C33.

Fehr, E. and Gächter, S. (2001): "Fairness and Retaliation: The Economics of Reciprocity", *Journal of Economic Perspectives*, 14, 159-181.

Fehr, E. and Schmidt, K. (1999): "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics*, 114, 817-868.

Fehr, E. and Schmidt, K. (2001): "Theories of Fairness and Reciprocity - Evidence and Economic Applications", forthcoming in: Dewatripont, M., Hansen, L. and Turnovsky, S. (Eds.): *Advances in Economics and Econometrics - 8th World Congress*, Econometric Society Monographs.

Fershtman, C. and Weiss, Y. (1998): "Why do we care what others think about us?", 133-151 in Ben-Ner, A. and Putterman, L. (eds.): *Economics, Values and Organization*, Cambridge University Press.

Frank, R. (1988): *Passions Within Reasons*, W.W. Norton & Co.

Guttman, J. (1999): "Self-enforcing Agreements and the Evolution of Preferences for Reciprocity", unpublished paper, Bar-Ilan University.

Guttman, J. M. (2000): "On the evolutionary stability of preferences for reciprocity", *European Journal of Political Economy*, 16, 31-50.

Güth, W., Schmittberger, R. and Schwarz, B. (1982): "An experimental analysis of ultimatum bargaining", *Journal of Economic Behavior and Organization*, 3, 367-388.

Güth, W. and Yaari, M.: (1992): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game". In Witt, Ulrich (ed.): *Explaining Process and Change - Approaches to Evolutionary Economics*, Ann Arbor, MI: University of Michigan Press.

Güth, W.: (1995): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives", *International Journal of Game Theory*, 24, 323-344.

Hofbauer, J. (1981): "On the occurrence of limit cycles in the Volterra-Lotka equation", *Nonlinear Analysis, Theory, Methods and Applications*, 5, p. 1003-1007.

Hofbauer, J. and Sigmund, K.: *Evolutionary Games and Replicator Dynamics*, Cambridge University Press, 1998.

Ledyard, John O., "Public Goods: A Survey of Experimental Research." In Kagel,

John, and Alvin Roth, eds., Handbook of Experimental Economics. Princeton: Princeton University Press, 1995, 111-194.

Ockenfels, P. (1993): "Cooperation in prisoner's dilemma", European Journal of Political Economy, 9, 567-579.

Ok, E. and Vega-Redondo, F. (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario", Journal of Economic Theory, 97, 231-254.

Robson, A. J. (1990): "Efficiency in evolutionary games: Darwin, Nash and the secret handshake", Journal of Theoretical Biology, 144, 379-396.

Roth, A. (1995): Bargaining experiments. In Kagel, J. and Roth, A. (eds): Handbook of Experimental Economics, 253-348, Princeton University Press.

Sethi, R. (1996): "Evolutionary stability and social norms", Journal of Economic Behavior and Organization, 29, 113-140.

Sethi, R. and Somanathan, E. (2001): "Preference Evolution and Reciprocity", Journal of Economic Theory, 97, 273-297.

Sethi, R. and Somanathan, E.: "Understanding Reciprocity", Journal of Economic Behavior and Organization, forthcoming.

Taylor, P.D. and Jonker, L.B.(1978): "Evolutionarily Stable Strategies and Game Dynamics", Mathematical Biosciences, 40, 145-156.

Vogt, C. (2000): "The evolution of cooperation in Prisoner's Dilemma with an endogenous learning mutant", Journal of Economic Behavior and Organization, 42, 347-373.

Weibull, J.W. (1995): Evolutionary Game Theory, MIT Press.

Zeeman, E.C. (1980): "Dynamics of Evolution of Animal Conflicts", Journal of Theoretical Biology, 89, 249-270.