Evolution of Reciprocal, Materialistic and Altruistic Preferences in an Environment of Prisoner's Dilemma Games

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

March 17, 2003

Abstract

The conventional assumption in economics, that individuals have 'materialistic' preferences, has been questioned by experimental evidence. In this paper we study the evolution of preferences when players are engaged in simultaneous and sequential move Prisoner's Dilemma games. There is, as long as each game is played with strictly positive probability, a unique asymptotically stable population where players with reciprocal, altruist and materialist preferences co-exist in the population. Our results provide some simple insights into the evolutionary foundations for the 'game of life' that may be regarded as causing much experimentally observed behavior.

Keywords: Social preferences; reciprocity; altruism; materialism; Prisoner's Dilemma game; simultaneous/sequential moves; multiple games; evolutionary stability.

JEL Classification: B41; C70; C72; D74; Z13.

1 Introduction

The conventional assumption in economics, that individuals are solely motivated by their material self-interest, has been questioned by experimental evidence. 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) and Cooper et. al. (1996)). 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) for an overview). Moreover, there are enough examples from everyday life: People give to charity and vote; they return lost wallets; they tip the waiter even when being sure to never visit the same restaurant again, and so on. And, on the negative side, a customer may refuse to buy a good from a seller in order to punish the seller for charging an 'unfair' high price; an employee who has been sacked may, on her last day at work, engage in acts of sabotage, in order to punish the employer (see Sobel (2001)).

In all these situations people behave in a way that does not maximize the monetary return, even when they do not have to worry about punishment or future interaction. 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/or 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' ('unkind') toward someone who is, or is expected to be, kind (unkind). Reciprocity is, in other words, *conditional* niceness. See e.g. Rabin (1993) and, for good surveys, see Fehr and Gächter (2000) and Sethi and Somanathan (2002). 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 final outcomes. 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.

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.

There is considerable empirical evidence indicating that a significant proportion of us have social preferences and that another significant proportion have materialistic preferences (see e.g. Fehr and Gächter (1998)). 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 proportions of individuals having different preferences. To do this, we let players with different preferences 'compete' against each other, in order to determine 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 use the one-shot Prisoner's Dilemma as the basic building block. The environment in which this game is played is varied along two dimensions: The *game form* (simultaneous-move, sequential or each being played with some probability, which we refer to as the *mixed* game) and *information* (perfect information or no information about other players' preferences). In the mixed version players are sometimes involved in simultaneous interaction and at other times in sequential interaction. We believe it is important to allow for such interaction in *multiple games*: The 'game of life' played 'out there' is not simply a simultaneous move game, or a sequential move game: There is a variety of game situations that we can find ourselves engaged in and they differ in many ways. In this paper we focus on the move structure: Due to unmodelled stochastic background factors, players sometimes end up playing a simultaneous game and other times they play a sequential game.[1]

We start by analyzing preference evolution in the simultaneous and the sequential game. We show that the preference dynamic is very different. We then turn our attention to the mixed game, where players are involved in simultaneous and sequential interaction with varying probabilities. Here preference selection under perfect information leads to a unique and asymptotically stable population. Three preferences types, namely materialist, reciprocator and altruist, co-exist in this population, with the proportions of each preference varying with the exogenous parameters. This co-existence result qualitatively captures the results from experiments. We obtain asymptotic stability for any probability distribution over the simultaneous and the sequential game, as long as each game is played with strictly positive probability. Allowing for multiple games thus strengthens the predictions of the model. The result also shows the importance of perturbing an analysis of a single game situation by introducing the possibility that players find themselves involved in other, nearby, situations.

Why do materialism, reciprocity and altruism co-exist? Intuitively, the reason is that neither preference type can dominate the other preference types: An all-altruistic population is too 'nice' to be evolutionarily stable; an all-materialistic population earns too low a material payoff; and, finally, an all-reciprocal population is not evolutionarily stable as long there is the tiniest chance that the players end up defecting against each other; this means that altruistic mutants can invade. The last has not, we believe, been seen in other models of preference evolution.[2] Moreover, there is a 'cyclical'

[1]For a different analysis of preference evolution in a variety of games, we refer the reader to Güth and Napel (2002).

[2]In the duopoly context analyzed in Bester and Güth (1998), however, an altruist may, under certain circumstances, be able to invade an all-materialist population. See also Bolle (2000) and

relationship between the three preferences: Against the reciprocal preference type, the altruist preference is optimal; and against the latter preference the materialist one is optima, against which reciprocity becomes optimal, and so on. In combination these two observations produce a unique and asymptotically stable population.

The second environmental variable that we change is information about fellow players' preferences. First, we conjecture that our results, derived under perfect information, will also hold for sufficiently precise information. If, however, players have zero information about other players' preferences, then social preferences have no impact on behavior: In any evolutionarily stable outcome, all players defect. We conjecture that the same will hold when players have sufficiently little information. 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 needs to be qualified: Certain environments are more conducive for social preferences than others. The 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 is necessary in order for two reciprocal players to perform well enough against each other and, moreover, it permits a Reciprocator to avoid being 'exploited' by Materialists. If, on the other hand, players have little or no information about their opponent's motivations, then reciprocity and altruism can not thrive in the population. See also Ok and Vega-Redondo (2001) for a similar result. 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. But, how is it that a player can deduce whether another person is, say, a reciprocal or materialist individual? Strictly speaking, of course, this is impossible: Individuals cannot 'see' inside other persons' heads. However, some 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.[3] 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. See e.g. Kandori (1992) for such a model. 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.

What is the relevance of our results for the experimental findings, mentioned earlier? In most experiments players do not have information about fellow players' preferences. Interaction is deliberately kept anonymous. Our results, emphasizing the role of information about preferences for reciprocity and altruism to survive, can therefore not directly explain to the experimental findings. What happens in the lab is pre-

Possajennikov (2000).

[3]We refer the reader to Brosig (2002) for an experiment examining these and other claims.

sumably that subjects' believe that there are sufficiently many reciprocal and altruist people out there. Given these (correct) beliefs subjects with reciprocal/altruistic preferences optimally cooperate. The question, therefore, once more, 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 players with other preferences would outperform them. The result is co-existence between players with different preferences. This is why some players, when seated in the lab in an anonymous setting, optimally co-operate, while others do not. We do not claim that our simple Game of Life, modeled as a mix of simultaneous and sequential Prisoner's Dilemma games, is an adequate representation of the real Game of Life. However, we still believe it gives an insight into what one might expect from a richer model.

There are several important differences between our model and the existing literature. First, we pay attention to how the Prisoner's Dilemma game is played: Do the players make choices simultaneously or sequentially? Or, are the players sometimes engaged in the simultaneous, and other times in the sequential game? We show this has a significant difference for the preferences that survive. For example, whereas materialistic players induce reciprocal players to defect under simultaneous interaction, it is reciprocal players who induce materialists to co-operate under sequential interaction (see Fehr and Schmidt (1999) for similar observations). Indeed, the evolutionarily stable preferences that emerge in the simultaneous and in the sequential game when played in isolation are qualitatively different from those that emerge when each game is played with arbitrary but strictly positive probabilities.

Second, rather than simply assuming, as is done in most of the existing literature, that two reciprocal individuals always manage to cooperate in the simultaneous Prisoner's Dilemma game (see e.g. Guttman (2000) and Ockenfels (1993)), we allow for the possibility that they sometimes defect. We believe this is more plausible, since the outcome where each reciprocal player defects is a strict Nash equilibrium and hence should not be a priori discarded. The possibility that two reciprocal players cannot always co-ordinate on their preferred equilibrium has the crucial implication that endogenous fluctuations occur and that players with altruistic preferences survive.

Third, the existing papers typically only allow for two kinds of preferences, namely the materialist and the reciprocal one (see e.g. Fershtman and Weiss (1998), Guttman (2000) and Ockenfels (1993)). We allow for altruism and give a complete analysis with these three preferences. We also derive some results where a fourth preference is admitted. We believe it is important to allow for as many preferences as possible: 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 (Sethi and Somanathan (2002) make the same point). We show that the models with only the Reciprocator and the Materialist preference types available may give too optimistic predictions about the occurrence of cooperation, due to their exclusion of players with altruistic preferences.

Fourth, and finally, we model social preferences in a way that differs somewhat from the one in the existing literature. There social preferences are modeled by positing a utility function that, in addition to a player's own material payoff, has fellow players' material payoffs as arguments (see e.g. Fehr and Schmidt (1999)). We, on the other hand, work directly with the underlying preference orderings. This allows us to study possible optimal behaviors that cannot always be captured by a specific functional form for the utility function. We elaborate on this in Section 2.2.2 below.

There are many other models of preference evolution, studying different games and using somewhat different modeling techniques. We refer the reader to Ely and Yilankaya (2001), Ok and Vega-Redondo (2001) and Sethi and Somanathan (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 'hardwired' to a certain behavior. We refer the reader to e.g. 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:

	C	D
C	1	b
D	a	0

where $a > 1 > 0 > b$, $(1/2)(a+b) < 1$ and 'C' and 'D' stand for 'Co-operate' and 'Defect', respectively. If both players care only about their own money payoffs, the unique outcome is (D, D), which, in terms of money, is worse than (C, C). In Section 3.1, we will assume this game is played simultaneously; in Section 3.2 we analyze it under sequential interaction. And then, in Section 4, a pair of players face each game with a strictly positive probability.

2.2 Preferences

Let (i, j), where $i, j = C, D$, denote the outcome where a player chooses i and the opponent chooses j. For reasons that will be discussed in Section 2.2.2 below, we will not represent these preferences by a utility function (possibly defined over the monetary consequences); we will instead consider the (pure) *best replies*, i.e., what a player will choose, given that the opponent plays C, and what will he choose, given the opponent plays D. We start by considering the following three *preference types*:

The ***Materialist*** (M) preference type: Choose D both if the opponent chooses C and if the opponent chooses D. That is, D is strictly dominant. The ***Reciprocator*** (R) preference type: Play C if the opponent chooses C and play D if the opponent

chooses D. The ***Altruist*** (A) preference type: Choose C both if the opponent chooses C and if the opponent chooses D. That is, C is strictly dominant.

In Section 5 below, we consider a 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 co-operation by the opponent is perceived as 'kind' and hence the person cooperates, too.[4] The Altruist is determined to try to establish a cooperative outcome and will always do her part, independently of what the opponent does. She can also be interpreted as wanting to maximize the sum of the players' monetary returns.[5]

However, we would like to stress that other interpretations of our preference types are possible. First, we could just as well have used the label 'Inequity averse' instead of 'Reciprocator'. This is because a player with sufficiently inequity averse preferences would do the same as the Reciprocator: 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, in this sense, just as 'social' as the other preference types. Of course, the following 'minimalist' interpretation is also 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). Our labels 'Materialist' etc., are then completely arbitrary and for convenience only.

2.2.2 Comparison with the Utility Function Approach to Modeling Social Preferences

A preference type as defined above, is a pair of pure best replies: What a player optimally chooses when the opponent chooses C (D). One may speculate what utility functions can generate these best replies. We therefore consider some of the functional forms proposed in the literature (see Fehr and Schmidt (2001) for an overview).

A popular utility function that has been used to model reciprocity was suggested in Fehr and Schmidt (1999). Suppose an outcome is realized that gives money payoffs (π_1, π_2) to players 1 and 2. Then player $i = 1, 2$ 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 $\alpha_i \geq 0$ and $\beta_i \geq 0$ represent player i's disutility from disadvantageous

[4] We refer the reader to Rabin (1993) for formal analysis of kindness.

[5] This requires that the parameters a and b satisfy $0 < a + b < 2$.

and advantageous inequality, respectively. These preferences are said to represent 'inequality aversion'. Depending on the parameter values different best replies for our PD-game emerge. If the opponent plays C, an individual plays C 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 our Reciprocator and the Materialist preference type, but never that of our Altruist preference type.

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

Whereas we can not get the behavior of our Altruist preference type from the specifications of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), we can get it, together with some of the other preference types, from a simple linear utility function (see e.g. Ledyard (1995)):

$$u_i(\pi_i, \pi_j) = \pi_i + \alpha_i \pi_j,$$

where α_i is the weight that person i assigns to person j's monetary payoff.[6] Some calculations reveal that a player with $\alpha_i < min\{(a-1)/(b-1), -b/a\}$ [$\alpha_i > max\{(a-1)/(b-1), -b/a\}$] has defection [co-operation] as a strictly dominant strategy, i.e., she is like our Materialist [Altruist] preference type. Furthermore, if $-b/a < (a-1)/(1-b)$ and $-b/a < \alpha_i < (a-1)/(1-b)$, we get the Reciprocator type, while if $-b/a > (a-1)/(1-b)$ and $(a-1)/(1-b) < \alpha_i < -b/a$, the player responds to cooperation with defection and responds to defection with cooperation. We will return to this, rather 'paradoxical', preference type in Section 5 below. Thus within this class of utility functions our we can get the Altruist, Materialist and Reciprocator preference types, but only for certain parameter values.

Finally, consider the utility function studied in Charness and Rabin (2001):

$$u_i(x_1, x_2) = (1-\gamma)\pi_i + \gamma\left[\delta\, min\{x_1, x_2\} + (1-\delta)[x_1 + x_2]\right],$$

where $\gamma \in [0, 1]$ and $\delta \in [0, 1]$. These preferences express a trade-off between one's own and both players' monetary payoff; the self-interest decreases with γ. Moreover,

[6] See Levine (1998) for an explicit modeling of these weights.

the concern for the groups's monetary payoff is a combination of a concern for the least advantaged and efficiency (the latter being weighted more heavily when δ is small). We verify that a player's best reply to a choice of C by the opponent is C when $\delta > \frac{1-2\gamma}{\gamma}$ (for which it is necessary that $\gamma > 1/3$) and otherwise D is optimal. Similarly, a best reply to D is C when $\delta < \frac{2\gamma-1}{2\gamma}$ (for which it is necessary that $\gamma > 1/2$). We can therefore get all our three A, R and M preference types for different (γ, δ) configurations. However, we cannot get the preference type for which a best reply to C (D) is D (C) (cf. Section 5 below).

We believe this shows that our approach, of postulating a set of preference types (i.e., various best reply combinations) is never more restrictive, and can be more general, than postulating a specific class of utility functions. Doing the latter may exclude some behavior that actually turns out to be 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 proportion 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 Pure Interaction

In this section players are engaged in either simultaneous or sequential interaction, but never both.

3.1 Simultaneous Interaction

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

Table 1: The money payoffs in the evolutionary game under simultaneous interaction. 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).[7] 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*

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

Assumption 1 implies that two Reciprocators perform better than two Materialists, but not as well as two Altruists. One may justify our assumption as follows: Suppose that when two R types meet, each individual plays C with probability λ and D with probability $1 - \lambda$ (λ could be interpreted as the probability assigned to C in the symmetric mixed Nash equilibrium). Then we have $0 < \pi_{RR} < 1$ whenever $0 < (1/2)(a + b) < 1$.

Proposition 1 *Consider the evolutionary game based on the simultaneous-move Prisoner's Dilemma game.*

(a). Any population where all players have the same preference is unstable.

(b). There is a unique interior equilibrium population, x^*. This equilibrium is a center, i.e., surrounded by periodic orbits. Thus, whenever the population is not initially exactly at x^*, the population frequencies of the three preference types fluctuate endlessly.

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

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

[7] There is also a symmetric and mixed Nash equilibrium, but we ignore it here.

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

Below we have given the dynamic for the case where $a = 2, b = -1$ and $\pi_{RR} = 1/2$.

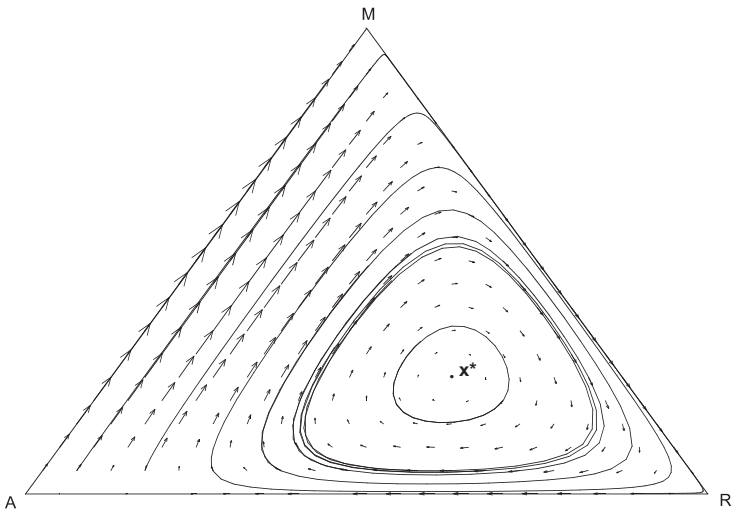

Figure 1: The phase diagram for the simultaneous move PD game; $a = 2$, $b = -1$ and $\pi_{RR} = 1/2$. Equilibrium proportions: $x_A^* = 1/4$, $x_R^* = 1/2$ and $x_M^* = 1/4$.

The vertex labeled i is the population where all players are of preference type i, where $i = A, R, M$. Any orbit circles around the equilibrium. The proportions of players with different preferences fluctuate endlessly.

Why do we get these never-ending fluctuations? As the proportion of R types increases, the proportion of A-types increases too and the proportion of M-types fall; however, once there are sufficiently many A-types, preference M gains foothold; this lowers the proportion of A and R-types, after which type R again gains territory, and so on. Moreover, 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 Table 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)). However, 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.[8]

Our results show that players with altruistic preferences perform equally well as those with materialistic or reciprocal preferences. Indeed, our 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.[9] There is a role for altruism whenever two reciprocators cannot perfectly establish full cooperation: Unlike a reciprocal player, an altruist induces a reciprocal player to co-operate with

[8] In Robson (1990) a similar argument is presented.

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

probability one.[10] Our results therefore show that the exclusion of altruism is not justified in the simultaneous-move game.

3.2 Sequential Interaction

We now assume that when two players are matched, one of them is randomly chosen to be first-mover. This player then chooses between 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.

Consider a player who, in the role of first-mover, faces a reciprocal second-mover. If the first-mover cooperates (defects), the second-mover cooperates (defects), too. Thus we must ask how the first-mover ranks the (C, C) outcome relative to the (D, D) 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 then 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

Table 2: The money payoffs in the evolutionary game under sequential interaction. A= Altruist; R = Reciprocator; M = Materialist.

We see that it is no longer true that the Altruist preference type is a unique best reply to the Reciprocator preference type. Moreover, the R preference type is a unique best reply to the M preference type. We get the following proposition:

Proposition 2 *Consider the evolutionary game for the Sequential Prisoner's Dilemma game.*

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

[10] If two Reciprocators could perfectly co-ordinate on co-operation, i.e., $\pi_{RR} = 1$, there would be no type M players in any stable population. See Poulsen (2002) for this analysis.

(b). A population is stable if it is composed exclusively of Reciprocators and Altruists and the proportion of Altruists, x_A, is sufficiently small: $x_A < 1/(2a-1)$.

The proof of this proposition is in the Appendix. We have illustrated the proposition below, where $a = 2$ and $b = -1$.

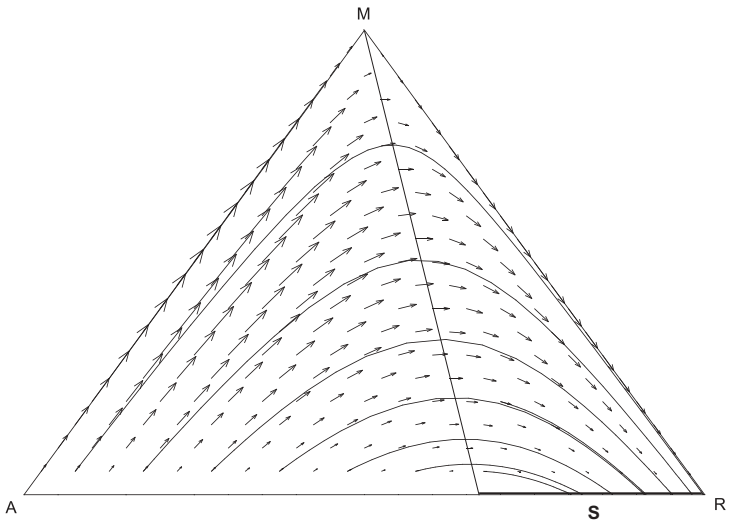

Figure 2: The phase diagram for the sequential PD game; $a = 2$ and $b = -1$.

The vertex M, where all players are Materialists, is unstable, as before (part (a)). The component labeled S consists of all the stable populations with Altruists and Materialists (part (b)). In our example, up to one-third of the players can be altruistic in a stable population. The vertical line connecting the M vertex with the endpoint of the S component gives those populations at which the proportion of M types is neither increasing or decreasing.

The important difference from simultaneous interaction is that (i). *Sequential interaction allows two Reciprocators to overcome their co-ordination problem* and (ii). a *Reciprocator outperforms a Materialist against a Materialist opponent.*[11] Thus the presence of reciprocally motivated players induces 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) and Fehr and Fischbacher (2002) for a discussion.

[11] 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 comes from the fact that when the Materialist is first-mover she optimally chooses to *cooperate* rather than to defect (cf. Assumption 2).

4 Mixed Interaction

In Section 3 interaction was either simultaneous or sequential. In this section we assume that when two players are matched, they engage in the simultaneous game with probability μ and with remaining probability $1-\mu$ interaction is sequential. We impose only $0 < \mu < 1$. We will refer to this game as the *mixed* PD game.

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	$\mu\pi_{RR} + 1 - \mu$	$(1-\mu)(1/2)$
M	a	$(1-\mu)(1/2)$	0

Table 3: The money payoffs in the mixed evolutionary game, where the simultaneous (sequential) game is played with probability μ $(1-\mu)$.

Let us recall that (i). in the simultaneous (sequential) game preference type A is the unique (alternative) best reply to preference type R, (ii), in both the simultaneous and the sequential game the M preference type is the unique best reply to preference type A and (iii). in the sequential (simultaneous) game preference type R is the unique (alternative) best reply to preference type M. This implies that for any $\mu \in (0,1)$ we get a *cyclical best reply structure*. The implication is given in the following proposition:

Proposition 3 *Consider the mixed evolutionary game, where individuals play the simultaneous Prisoner's Dilemma game with probability μ and play the sequential game with probability $1-\mu$. There is for each $\mu \in (0,1)$ a unique interior equilibrium, $y^* = [y_A^*, y_R^*, y_M^*]$ and it is globally asymptotically stable.*

$$y_A^* = \frac{\mu^2 + 2b(2\pi_{RR} - 1)\mu + 2b - 1}{E}, \tag{2}$$

$$y_R^* = \frac{2(a-1)[2b - 1 + \mu]}{E}, \tag{3}$$

$$y_M^* = \frac{4\mu(1-a)(1-\pi_{RR})}{E}, \tag{4}$$

and with $E = 4\mu\pi_{RR}(a+b-1) + (2b-\mu-1)(2a-1-\mu)$.

The proof is in the Appendix. An illustration is provided below, again with $a = 2$, $b = -1$ and $\pi_{RR} = 1/2$, and with each game being equally likely to be played: $\mu = 1/2$.

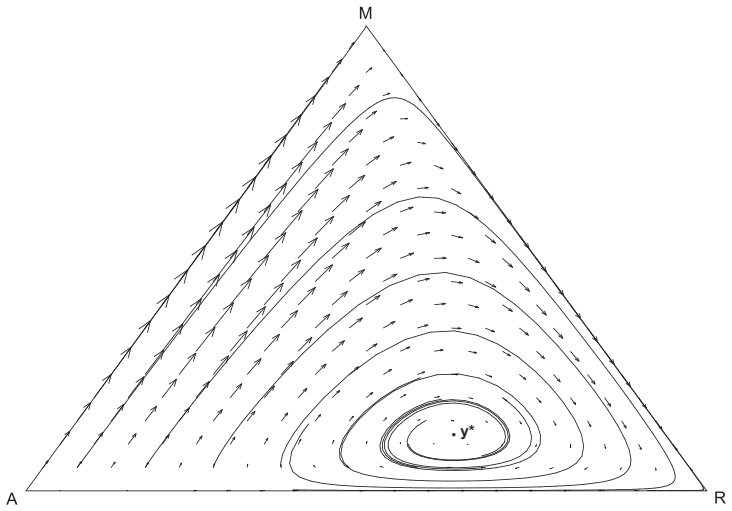

Figure 3: Phase diagram for the mixed PD game. Parameter values: $a = 2$, $b = -1$, $\pi_{RR} = 1/2$ and $\mu = 1/2$. Equilibrium proportions: $x_A^* = 11/35$, $x_R^* = 20/35$ and $x_M^* = 4/35$.

Evolutionary selection for the mixed game gives a unique prediction of population behavior for the mixed game for any value of $\mu \in (0, 1)$. Any initial population, where all three types are present, will over time evolve to the preference profile y^*. It is the cyclical best replies mentioned above that creates the spiraling convergence to the equilibrium (see also Hofbauer and Sigmund (1998)). Thus, by allowing for multiple games being played, we obtain a sharper prediction than in either the simultaneous or the sequential game. Our result qualitatively mimics the finding from the experimental literature that there are several personality types in the population and these are materialistic, reciprocal and altruist.

5 Including the 'Paradoxical' Preference Type

Until now we ignored the following preference type:

The ***Paradoxical*** (P) preference type: Play C if the opponent plays D and play D if the opponent plays C.

In this section we consider evolutionary stability when players may also evolve this type of preference. We proceed straight to the mixed game. Let π_{ij}, where $i, j = A, R, M, P$, denote the money payoff to a player of preference type i when matched with an opponent of preference type j in the simultaneous game. The matrix below contains all the payoffs for the evolutionary game.

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

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

Let us briefly 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 prefers the outcome (D, C) over outcome (C, D), i.e., chooses D. When the P-type is first-mover, Assumption 2 (Section 3.2) 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[12], 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 $(1/2)\pi_{PP} + (1/2)[(1/2)a + (1/2)b]$.

Looking at the payoffs in the matrix reveals that the M type performs strictly better, or at least as well, as the P type against the A, M and P types. Moreover, type M performs better against an R type in the sequential game. This follows since the P type's inclination to choose D (C) if the opponent plays C (D), coupled with our assumption that the R type prefers (D, C) over (C, D), means that the P type co-operates and the R type defects. The P type's preferences lead her to being exploited by an R type in the sequential game. The M-type, on the other hand, realizes the better (D, D) outcome when matched with an R-type and the former is second-mover. However, the P type performs strictly better than the M type against an R type in the simultaneous contest. This gives us the following proposition:

Proposition 4 *Suppose it is sufficiently likely that interaction is sequential:* $\mu < \frac{-b}{2\pi_{PR}-b}$. *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.

The condition in the proposition is likely to hold whenever the P type performs badly against the R type in the simultaneous-move setting (such that π_{PR} is small).

[12]There are also two asymmetric pure Nash equilibria, but we ignore them here.

This, in turn, will be the case whenever the R-type's preferences lead her to behave 'aggressively' against the P type, i.e., to be very likely to play defect.

Part (b) implies that we can effectively ignore the P preference type 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 the these three strategies were available from the beginning. In this case all our results from Sections 3 and 4 hold.

6 The Case of No Information about Preferences

In the previous analysis subjective payoffs were common knowledge. It was as if players had their preference types written on their foreheads. We conjecture that our results continue to hold as long as information is *sufficiently* accurate, or as long as it is not too costly to acquire such information.[13] However, let us now consider the polar opposite to perfect information: 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. All players face 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. The following result holds no matter whether interaction is simultaneous, sequential or mixed:

Proposition 5 *Consider the simultaneous, the sequential or the mixed evolutionary game when players do not know their opponents' preferences. Then: In any stable population there are no Altruists and all players defect.*

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'.[14]

This proposition underlines the importance of personal communication for reciprocity, and altruism, to be successful. Reciprocal players must be able to signal, or communicate, what value system they have, in order to establish cooperative outcomes and in order to avoid co-operating with 'bad' players. In a *completely* anonymous world, cooperation is not possible.

[13] For a formal analysis, see Güth (1995) and Güth and Kliemt (1994).

[14] See also Ok and Vega-Redondo (2001).

The assumption of no information seems just as unrealistic as one of perfect information. In the real world, individuals form opinions about fellow individuals' preferences based on information about observables such as income and skin color. The future challenge, we believe, is to explore this area 'intermediate' between perfect and no information.

7 Conclusion

In this paper we analyzed 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. In particular, we assumed that players were involved in several game situations with varying probabilities. When players have information about other players' preferences, a unique asymptotically stable distribution of preferences emerged over time. Here reciprocal, materialist and altruist individuals live side-by-side. We believe our results may 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, 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. Table 1): $\pi(A,x) = 1 - x_M + x_M b$, $\pi(R,x) = x_A + x_R \pi$ and $\pi(M,x) = x_A a$. 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:

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

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

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

We next 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 the Replicator Dynamic: If (p,q) is a fixed point for the Lotka-Volterra dynamic, then

$$x^* = (x_A^*, x_R^*, x_M^*) = (1/(1+p+q), p/(1+p+q), q/(1+p+q)) \tag{5}$$

is a fixed point for the Replicator Dynamic. Moreover, results about the stability of one system will hold for the other system (Hofbauer (1981)). We refer the reader to Bomze (1983) for details. See also Hofbauer and Sigmund (1998).

We may, instead of the matrix in Table 1, study the equivalent matrix, obtained by deleting the number 1 (1) [b] from all entries the first (second) [third] column:

	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 if 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), where $p > 0$ and $q > 0$ and

$$p = \frac{-b(a-1)}{-\pi b}$$

$$q = \frac{(1-\pi)(a-1)}{-\pi b}.$$

We may verify that when these expressions are used in (5), we get exactly the solutions above and in the main text.

Bomze shows furthermore 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^*, given in (5) above, is a center for the Replicator Dynamic. Finally, in order to verify that $0 < x_i^* < 1$ for $i = A, R, M$, we may use the fact that $p > 0$, $q > 0$ and (5). ■

Proof of Proposition 2: 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 Dynamic[15], 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 3:

The equations giving the expected payoffs are now $\pi(A,x) = x_A + x_R - x_M b$, $\pi(R,x) = x_A + x_R[\mu\pi + 1 - \mu] + x_M[(1-\mu)(1/2)]$ and $\pi(M,x) = x_A a + x_R[(1/2)(1-\mu)]$. Solving these equations yields

$$y_A^* = \frac{-1 + \mu^2 - 2\,b\mu + 2\,b + 4\,b\mu\,\pi}{1 - 4\,\mu\,\pi - 2\,\mu\,a - 2\,b\mu + 4\,ab + 4\,b\mu\,\pi + 4\,\mu\,\pi\,a + \mu^2 - 2\,b + 2\,\mu - 2\,a}$$

$$y_R^* = 2\,\frac{1 - a - 2\,b + 2\,ab + \mu\,a - \mu}{1 - 4\,\mu\,\pi - 2\,\mu\,a - 2\,b\mu + 4\,ab + 4\,b\mu\,\pi + 4\,\mu\,\pi\,a + \mu^2 - 2\,b + 2\,\mu - 2\,a}$$

$$y_M^* = 4\,\frac{\mu\,(-\pi - a + \pi\,a + 1)}{1 - 4\,\mu\,\pi - 2\,\mu\,a - 2\,b\mu + 4\,ab + 4\,b\mu\,\pi + 4\,\mu\,\pi\,a + \mu^2 - 2\,b + 2\,\mu - 2\,a}$$

Simplifying these expressions gives those in the main text.

Consider now the dynamic stability of y^*. Again, subtracting 1 (1) [b] from the first (second) [third] column from the matrix in Section 4 gives us the following equivalent matrix:

	A	R	M
A	0	0	0
R	0	$\mu\pi - \mu$	$(1/2)(1-\mu) - b$
M	$a-1$	$-(1/2)(1+\mu)$	$-b$

Using the same notation as earlier, we compute $\beta\theta - \gamma\epsilon = (1/2)b[\mu - 1 - 2\mu\pi] + (1/4)(1-\mu)(1+\mu)$, $\alpha\epsilon - \beta\delta = \mu(1-\pi)(a-1)$ and $\gamma\delta - \alpha\theta = (a-1)[(1/2)(1-\mu) - b]$. It is straightforward to verify that all three expressions are strictly positive. Thus we may, once more, use the results in Bomze (1983) to conclude that there is a unique fixed point (p,q) for the Lotka-Volterra dynamic, where $p > 0$ and $q > 0$ and

$$p = \frac{(a-1)[(1/2)(1-\mu) - b]}{(1/2)b[\mu - 1 - 2\mu\pi] + (1/4)(1-\mu)(1+\mu)}$$

[15] We refer the reader to e.g. Weibull (1995).

$$q = \frac{\mu(a-1)(1-\pi)}{(1/2)b[\mu - 1 - 2\mu\pi] + (1/4)(1-\mu)(1+\mu)}.$$

Again, using the relationship $y^* = (y_A^*, y_R^*, y_M^*) = \left(\frac{1}{1+p+q}, \frac{p}{1+p+q}, \frac{q}{1+p+q}\right)$ gives our expressions in the main text. Furthermore, we compute

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

The numerator is strictly negative for any $\mu \in (0,1)$. Since the denominator is strictly positive, we may conclude that $\beta p + \theta q < 0$. Thus (p,q) is asymptotically stable for the Lotka-Volterra dynamic. This then implies that y^* is asymptotically stable for the Replicator Dynamic. The proof that $0 < y_i^* < 1$ follows from using the relationship between p, q and y^* given earlier and the fact that $p > 0$ and $q > 0$. ■

Proof of Proposition 4:

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 (see Weibull (1995), Proposition 3.2). 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. ■

Proof of Proposition 5:

It is not difficult to see that there can be no Altruists in any stable population. For the Altruists always cooperate and 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. ■

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.

Bester, H. and Güth, W. (1998): "Is altruism evolutionarily stable?", Journal of Economic Behavior and Organization, 34, 193-209.

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.

Bolle, F. (2000): "Is altruism evolutionarily stable? And envy and malevolence?", Journal of Economic Behavior and Organization, 42, 131-133.

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.

Brosig, J. (2002): "Identifying cooperative behavior: some experimental results in a prisoner's dilemma game", Journal of Economic Behavior and Organization, 47, 275-290.

Charness, G. and Rabin, M. (2001): "Understanding Social Preferences with Simple Tests", forthcoming in Quarterly Journal of Economics.

Cooper, R., DeJong, D., Forsythe, R. and Ross, J. (1996): "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games", Games and Economic Behavior, 12, 187-218.

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.

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 Gächter, S. (1998): "Reciprocity and economics: The economic implications of *Homo Reciprocans*", European Economic Review, 42, 845-859.

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

Güth, W. and Kliemt, H (1994): "Competition or co-operation - On the evolutionary economics of trust, exploitation and moral attitudes", Metroeconomica, 45, 155 - 187

Güth, W. and Napel, S. (2002): "Inequality Aversion in a Variety of Games - An Indirect Evolutionary Approach", Working paper 23-2002, Max Planck Institute for Research into Economic Systems.

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.

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.

Kandori, M. (1992): "Social Norms and Community Enforcement", Review of Economic Studies, 59, 63-80.

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.

Levine, D. (1998): "Modeling Altruism and Spitefulness in Experiments", Review of Economic Dynamics, 1, 593-622.

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.

Poulsen, A. (2002): "On Prisoner's Dilemma Payoffs and the Evolution of Cooperative Preferences", unpublished manuscript, Department of Economics, Aarhus School of Business.

Possajennikov, A. (2000): "On the evolutionary stability of altruistic and spiteful preferences", Journal of Economic Behavior and Organization, 42, 125-129.

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

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): Hand-

book 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. (2002): "Understanding Reciprocity", Journal of Economic Behavior and Organization, 50, 1-27.

Sobel, J. (2001): "Interdependent Preferences and Reciprocity", working paper, Department of Economics, University of California, San Diego.

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.