CBESS Discussion Paper 09-02

Altruism and Welfare When Preferences Are Endogenous

by Anders Poulsen* and Odile Poulsen

*School of Economics, University of East Anglia

Abstract
We study whether an altruistic preference can survive in competition with other preferences and investigate the relationship between the equilibrium proportion of altruism and equilibrium material and subjective welfare. Altruism survives whenever preferences are sufficiently observable. Altruism can co-exist with reciprocal and materialistic preferences. Any policy that increases the equilibrium proportion of altruism raises economic prosperity but can reduce some people's subjective equilibrium welfare. Some of the policies that increase the equilibrium proportion of altruism are, at first sight at least, counter-intuitive. There can be a non-monotonic relationship between the degree of anonymity of interaction in society (the probability that an individual knows other people's preferences) in society and welfare.

JEL classification codes
C70, D60, D64

Keywords
Altruism; endogenous preferences; material and subjective welfare; indirect evolutionary approach.

University of East Anglia

CBESS

Centre for Behavioural and Experimental Social Science

CBESS
University of East Anglia
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/ssf/cbess

Altruism and Welfare When Preferences Are Endogenous

Anders Poulsen* and Odile Poulsen
School of Economics, University of East Anglia
Norwich NR4 7TJ, United Kingdom
E-mail: a.poulsen@uea.ac.uk and o.poulsen@uea.ac.uk

October 13, 2009

Abstract

We study whether an altruistic preference can survive in competition with other preferences and investigate the relationship between the equilibrium proportion of altruism and equilibrium material and subjective welfare. Altruism survives whenever preferences are sufficiently observable. Altruism can co-exist with reciprocal and materialistic preferences. Any policy that increases the equilibrium proportion of altruism raises economic prosperity but can reduce some people's subjective equilibrium welfare. Some of the policies that increase the equilibrium proportion of altruism are, at first sight at least, counter-intuitive. There can be a non-monotonic relationship between the degree of anonymity of interaction in society (the probability that an individual knows other people's preferences) in society and welfare.

Keywords: Altruism; endogenous preferences; material and subjective welfare; indirect evolutionary approach.

JEL Classification: C70; D60; D64.

1 Introduction

A definition of altruism is: "The principle or practice of unselfish concern for or devotion to the welfare of others."(The Random House Webster's Unabridged Dictionary (1997)). Altruism is also often defined as the act of giving up material resources in order to increase those of other individuals (see, for example, Cambridge Advanced Learner's Dictionary). It has been a challenge for social science, and for biology, to explain why such behavior is observed in various situations, such as charity, redistribution, self-sacrifice, and several other kinds of helping behavior. How can altruistic behavior in economic situations be rationalized when it, apparently, would

*Corresponding author. Anders Poulsen thanks the Danish Social Science Research Council for financial assistance (Grant 212.2269.01).

perform worse than behavior directed at maximizing material returns? The classical economists, such as Bentham, Edgeworth, Hume, Mill, and Smith, wrote about altruism; see the survey in Collard (1978). Among more recent contributions are Andreoni (1990), Becker (1974, 1976), Bergstrom (1989), Bernheim and Stark (1988), Bernheim, Shleifer and Summers (1985), Bruce and Waldman (1990), Buchanan (1977), Hirshleifer (1977), Hochman and Rodgers (1969), Rotemberg (1994) and Stark (1989), and several papers using an evolutionary approach. These are described in Section 6.

Most of the work by the classical economists is concerned with the question of whether individuals are, or ought to be, altruistic; it is not discussed why and how altruistic behavior emerges. The more recent papers mentioned above analyze the behavioral and welfare effects of altruism. Many important insights on the effects of altruism have been obtained. However, these contributions typically focus on interaction between two persons, between persons within a single family, or on behavior within a small group only. Also, it is again not (with some exceptions, such as Rotemberg, 1994) explained *why* some individuals are altruists; this is simply taken as given. Moreover, it is (implicitly) assumed that individuals know each other's degrees of altruism (preferences are common knowledge); this may be appropriate within a family, but not in a large society where interaction is random and anonymous.

In this paper we consider the emergence and material and subjective welfare implications of altruism at the societal level. We pose the following questions:

- **1:** When is there a societal *equilibrium preference distribution* where *endogenously* some individuals are altruists?
- **2:** How do various policies affect the equilibrium proportion of altruism?
- **3:** Is altruism 'good' for society? That is, does any exogenous policy that increases (reduces) the equilibrium proportion of altruism also increase (reduce) individuals' material or subjective equilibrium welfare?
- **4:** How does the equilibrium proportion of altruism depend on the degree of anonymity of interaction in society?

In our model interaction in society is represented by the Prisoner's Dilemma game. This simple social dilemma seems appropriate for our purposes: to provide a sufficiently serious test of the the survival ability of altruism we should examine its performance in an environment that is quite harsh and hostile. In the model individuals can differ in their preferences. We define an altruist as someone whose preferences lead him to optimally choose the co-operative action no matter what the opponent does; this definition of altruism is admittedly very stark, but also very simple. The distribution of preferences in society is endogenized by postulating a dynamic process by which the proportion of materially successful preferences increases at the expense of those performing less well. Only preferences that generate economically viable behavior survive over time. Thus, rather than simply assuming that some people are altruistic, we ask whether altruism can pass the test of economic survival. This analytical approach is known as the 'indirect evolutionary approach' and it was pioneered in Güth and Yaari (1992), and has since then shown its usefulness in several applications. See, for example, Berninghaus, Korth, and Napel (2007), Bester and Güth

(1998), Güth (1995), Güth and Kliemt (1998), Güth and Napel (2006), Königstein and Müller (2000), and Poulsen and Poulsen (2006).

We obtain the following answers to our questions:

- **1:** Altruism survives in society when the altruist's preferences are sufficiently observable by other people.

An altruist always co-operates, so he performs well against another altruist. Moreover, whenever the altruist's preferences are observed by other people he performs just as well against individuals with a preference for *reciprocity* (see Fehr and Gächter, 2002), since such individuals reciprocate the altruist's co-operation. However, an altruist is exploited by materialistic individuals who defect. When preferences are sufficiently observable, there is a preference distribution where the gains balance the costs, and then altruism survives in the endogenous preference distribution. In fact, there can be an endogenous preference distribution where altruism survives together with reciprocity and materialism. Analysing this endogenous preference distribution allows us to obtain answers to Questions 2 and 3.

- **2:** Policies that, at first sight, seem conducive to altruism can have the opposite effect. More precisely, a policy that *decreases* the monetary loss to an altruist from interacting with a materialistic person can *decrease* the equilibrium proportion of altruism.

In the short run, for a given preference distribution, a policy that improves the altruists' money earnings triggers an adjustment in the underlying preference distribution because the material returns to altruism at the original preference distribution now exceed the returns to the other preference types. But since materialistic persons are those that benefit the most from interacting with altruists, the net result in the longer run is an inflow of materialism, at the expense of altruism.

As an example we can think of a government policy that raises compensation to citizens who become victims of opportunistic or criminal behavior (and does not affect the returns from criminal behavior). In the short run, the policy raises the material returns to law-abiding citizens; there will thus be a tendency for people to abandon criminal behavior and become law-abiding citizens. However, since criminal persons are those who benefit most from a decrease in crime (since there are now more lab-abiding people that can be preyed on), the net effect of the policy is to *increase* criminal behavior. A policy that decreases a person's direct gains from criminal behavior (by, say, increasing the probability that the criminal is caught by the police or by increasing punishment) will, on the other hand, increase the equilibrium proportion of altruism. The difference between the two policies is that the first policy does not affect criminals' returns directly, whereas the second policy does.

- **3:** An increase in the equilibrium proportion of altruism always increases material welfare. However, if some individuals are 'too' altruistic, their subjective equilibrium welfare can fall.

A policy that raises the equilibrium proportion of altruism makes all individuals materially better off and this makes them subjectively better off *per se*. However, there is another effect: Individuals who care about other individuals' material welfare are affected by changes in the equilibrium proportions of the latter individuals. If people in the first group are sufficiently altruistic these fluctuations have a negative effect on subjective welfare, and it outweighs the positive effect on subjective welfare that is due to the increase in their personal material welfare. Furthermore, and importantly, we emphasise that these results are consistent with individuals never caring more for other individuals' money earnings than for their own. In other words, our results do not require that people are 'excessively' altruistic.

Finally, regarding the last question our main result is that

- **4:** A reduction in the degree of anonymity of interaction, which makes it more likely that people's preferences are observed by others, reduces the equilibrium proportion of altruism if there is already sufficiently much information available. There is a unique degree of anonymity at which material equilibrium welfare is maximized and this optimal degree is *not* perfect observability.

Increasing information about preferences is good for altruism, since it makes it more likely that reciprocal individuals reciprocate the altruist's co-operation. It thus increases the equilibrium proportion of altruism. If, however, the amount of available information increases beyond a certain limit, materialists enter population. From that point and onwards an increase in information increases materialism at the expense of altruism and is thus not desirable. Thus the 'best' degree of anonymity is the one that is high enough to allow altruism to survive but at the same time low enough to prevent materialists from entering the population.

The rest of the paper is organized as follows: Section 2 sets up the basic model, describing the strategic situation, preferences and how the distribution of preferences changes over time. In Section 3 we study when altruism is present in the endogenous preference distribution (Question 1), assuming initially that preferences are perfectly observable. Section 4 analyzes how different policies affect the equilibrium proportion of altruism (Question 2) and the welfare implications of altruism (Question 3). To study Question 4, Section 5 relaxes the assumption that preferences are common knowledge. We show that altruism survives as long as preferences are *sufficiently* observable. In Section 6 we relate our paper to the existing evolutionary literature. Section 7 summarizes our findings. All proofs are in the Appendix.

2 The Model

2.1 The Strategic Environment

The matrix below shows the money payoffs in the Prisoner's Dilemma (PD) game:

	C	D
C	r	s
D	t	p

Each person has two actions, Co-operate (C) and Defect (D). 'r' stands for 'reward', 's' for 'sucker', 't' for 'temptation', and 'p' for 'punishment', where $t > r > p > s$. There is a single large population of individuals. Time is continuous and at every time instant individuals are randomly matched in pairs and play the PD game, where, as already mentioned, the payoffs s, p, r and t are interpreted as *money* payoffs.

We assume that actions in the PD game are strategic complements, that is, the 'marginal' return from co-operation is greater when the opponent co-operates than when he defects:

$$r - t > s - p. \tag{1}$$

The role of this assumption is explained in the next section.

2.2 Preferences

Let $\pi(i, j)$ denote the money payoff to an individual when he chooses i and the opponent chooses j, where $i, j = C, D$. We assume that an individual's preferences can be represented by the following payoff function:

$$u(i, j) = \alpha\pi(i, j) + (1 - \alpha)\pi(j, i), \tag{2}$$

The same payoff function is used in Bester and Güth (1998); it was first studied by Edgeworth (1881). The parameter $\alpha \geq 0$ measures the weight assigned to the individual's own monetary payoff. We refer to α as the degree of material self-interest, and to $1 - \alpha$ as the degree of altruism. Individuals can differ in their α value and this give rise to different preferences, and hence different behavior. In order to avoid the potential criticism that our results depend on 'excessive altruism', we impose the restriction that individuals never care more about their opponent's money payoffs than about their own:

$$\alpha \geq 1/2. \tag{3}$$

We impose (3) only to ensure an economically plausible interpretation of our results. It is not imposed in order to avoid the so-called 'After-you' problem, that can occur when individuals are excessively concerned about opponent's (subjective) payoffs; see Collard (1978). This problem would not occur in our set-up even if (3) were dropped.

When $\alpha < (r-s)/(t-s)$ an individual whose opponent plays C optimally plays C. Similarly, the optimal reply to an opponent who plays D is D whenever $\alpha > (t-p)/(t-s)$. Strategic complementarity, as defined in (1), implies $(t-p)/(t-s) < (r-s)/(t-s)$, so an individual whose α parameter satisfies

$$\frac{t-p}{t-s} < \alpha < \frac{r-s}{t-s} \tag{4}$$

plays C if the opponent plays C, and plays D if the opponent plays D. An individual whose parameter satisfies

$$\alpha > max\left\{\frac{r-s}{t-s}, \frac{t-p}{t-s}\right\} = \frac{r-s}{t-s} \tag{5}$$

has D as a strictly dominant action. And, finally, an individual with parameter

$$\alpha < min\left\{\frac{r-s}{t-s}, \frac{t-p}{t-s}\right\} = \frac{t-p}{t-s} \tag{6}$$

strictly prefers to co-operate. (6) is consistent with (3) when $(t-p)/(t-s) > 1/2$, or

$$(1/2)(s+t) > p. \tag{7}$$

We assume this holds.

We call an individual whose α parameter satisfies the inequality (4) a *Reciprocator* (R); an individual whose parameter satisfies the inequality (5) is called a *Materialist* (M); and, finally, an individual whose degree of altruism satisfies the inequality (6) is an *Altruist* (A). It turns out that individuals whose parameters satisfy the same of the three inequalities (4)–(6) above behaves identically against any opponent. Thus there is no loss of generality in restricting attention to only three values of the α parameter, each satisfying one of the three inequalities.

We initially assume that preferences are common knowledge. Our results in Sections 3 and 4 are obtained under this assumption. In Section 5 preferences are only imperfectly observed.

2.3 Optimal Behavior for Given Preferences

A Materialist always defects and an Altruist always co-operates, no matter the opponent's preferences. When two Reciprocators meet they are involved in a co-ordination game with three Nash equilibria: (C, C), (D, D), and a symmetric mixed Nash equilibrium. This is sometimes called an 'Assurance game' (see Sen (1967)). Two Reciprocators thus face an equilibrium selection problem. We make the following assumption:

Assumption 1 *Two Reciprocators select the* (C, C) *Nash equilibrium with probability* λ *and select the* (D, D) *Nash equilibrium with probability* $1-\lambda$*, where* $\lambda \in [0, 1]$.

The same assumption is made in Fershtman and Weiss (1998). If $\lambda = 1$ there is perfect co-operation, while $\lambda = 0$ means there is only defection. If $0 < \lambda < 1$ there is *some*, but not full, co-operation. We believe the case $0 < \lambda < 1$ is a quite plausible. It can be interpreted as a reduced form of the interaction between reciprocal individuals when they sometimes, due to misunderstandings, various kinds of noise, or due to a lack of trust, fail to select the cooperative equilibrium with probability one. As Gambetta (1988) writes,

> The first player's anticipation of the second's defection may be based simply on the belief that the second player is unconditionally uncooperative.

> But, more tragically, it may also be based on the fear that the second player will not trust *him* to cooperate, and will defect as a direct result of this lack of trust. Thus the outcome converges on a sub-optimal equilibrium, *even if* both players might have been *conditionally* predisposed to cooperate (...). (p. 216, italics in original).

2.4 Evolution of Preferences

Denote the expected money payoff that a Reciprocator gets from interacting with another Reciprocator as μ, where

$$\mu = \lambda r + (1 - \lambda)p. \tag{8}$$

In the matrix below entry (i, j) is the expected money payoff to an individual with preference i when matched with another individual with preference j, where $i, j = R, M, A$ (R=Reciprocator, M=Materialist, A=Altruist). For example, entry (R, M) contains the payoff p because each individual plays D; similarly, when a Reciprocator and an Altruist meet, each individual plays C.

	R	M	A
R	μ	p	r
M	p	p	t
A	r	s	r

Table 1: The expected money payoffs in the evolutionary game.

Time is continuous and at any time instant t the state of the population is a *preference distribution*,

$$x(t) = (x_R(t), x_M(t), x_A(t)),$$

where (we suppress the time variable t from now on) $x_i \geq 0$ and $\sum_i x_i = 1$ and $i = R, M, A$. The number x_i is the proportion of individuals of preference i (at time t). The distribution of preferences change over time, in response to selection pressures based on *money* earnings. Precisely, denote by $\pi(i, x)$ the expected money earnings of preference type $i = R, M, A$, when the preference distribution is $x = (x_R, x_M, x_A)$. Let also $\pi(x, x)$ denote the average money earnings in the population x: $\pi(x, x) = \sum_i x_i \pi(i, x)$. The evolution of the preference distribution is governed by the Replicator Dynamic:

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

where $i = R, M, A$ and where $\dot{x}_i$ is the time derivative of x_i. The growth rate of individuals with preference i is positive if and only if such individuals make above-average money earnings, Thus only money matters for selection of preferences, not

subjective payoffs. The Replicator dynamic is due to Taylor and Jonker (1978); we refer the reader to for example Weibull (1995) for a description of this, and related, selection dynamics.

Our interpretation of this dynamic process is that of an economic or social selection mechanism according to which *materially* successful preferences are imitated. This process may work in many several ways, see for example the discussion in Güth and Kliemt (1998) and in Witt (1991)). Parents choose which preference to endow their children with. These parents choose a preference that they find have been successful, and 'successful' means high money earnings. Thus, even altruistic parents decide to instill a materialistic (reciprocal) preference in their children if the parents have observed that materialistic (reciprocal) individuals in the current period made more money than altruists.

3 Question 1: Can an Altruistic Preference Survive?

In this section we show how the survival of altruism depends on how well reciprocal individuals perform against each other.

3.1 Reciprocators can Establish Some But Not Full Co-operation

We start with the situation where two reciprocal individuals can establish *some*, but not perfect, co-operation: $0 < \lambda < 1$. If $\lambda < 1$ then $\mu < r$. Thus the Altruist preference type is a unique best reply to the Reciprocator preference: An Altruist induces a Reciprocator to co-operate with probability one with the Altruist, and hence the Reciprocator treats the Altruist better than he treats any other individual. We refer to this as a *strategic effect* (Schelling, 1978) of the Altruist's preferences on opponents' behavior. In the evolutionary model, this implies that an Altruist preference can invade an all-Reciprocator preference distribution. Moreover, $\lambda > 0$ implies $\mu > p$, and so the Reciprocator preference invades the all-Materialist preference distribution. We then have the following result:

Proposition 1 *Let $0 < \lambda < 1$. There is a unique stable equilibrium, x^*, for the Replicator Dynamic, and all three preferences are present: $x_i^* > 0$ for all $i = R, M, A$. The equilibrium is a center, that is, x^* is surrounded by closed orbits.*

$$x^* = (x_R^*, x_M^*, x_A^*) = \left(\frac{(p-s)(t-r)}{D}, \frac{(r-\mu)(t-r)}{D}, \frac{(\mu-p)(p-s)}{D} \right), \qquad (9)$$

where $D = (\mu - p - r)[p + r - (s + t)] + pr - ts$, and where μ is given in (8) above.

Proof: See the Appendix.

Whenever the initial preference distribution differs from x^*, the proportions of the three preferences permanently fluctuate and the fluctuations neither die out nor explode. Intuitively, whenever a large proportion of individuals carry one of the three preferences, individuals with another preference are more successful and displace the former. This gives rise to the perpetual cycles in the preference proportions.

3.2 When Reciprocators Either Fully Co-operate or Fully Defect

In this section we assume two Reciprocators select either the (C, C) or the (D, D) Nash equilibrium. This corresponds to $\lambda = 1$ or $\lambda = 0$, respectively.

Proposition 2 *(a). Suppose two Reciprocators play the (C, C) Nash equilibrium ($\lambda = 1$). Then altruism can survive with reciprocity as long as there are sufficiently few Altruists. Precisely, any population with Reciprocators and Altruists only, and where the proportion of Altruists is below $(r-p)/(t-p)$, is a Neutrally Stable Strategy (NSS) and hence stable for the Replicator Dynamic.*

(b). Suppose two Reciprocators play the (D, D) Nash equilibrium ($\lambda = 0$). Then altruism does not survive. More precisely, there is no stable preference distribution with Altruists.

Proof: Please see the Appendix.

We refer the reader to Weibull (1995) for the NSS concept, due to Maynard-Smith (1982). Part (a) follows from the fact that when two Reciprocators cooperate with probability one, an Altruist no longer outperforms the Reciprocator against another Reciprocators; instead the Altruist and the Reciprocator perform equally well. An Altruist can therefore no longer displace the all-Reciprocator population, but can slowly enter the population (this is known as 'evolutionary drift'; see Sethi and Somanathan, 2003). Indeed, in any population composed only of reciprocal and altruistic individuals, all individuals co-operate with each other. Thus the Reciprocators and Altruists perform equally well in the population. Any such population is stable as long as a materialistic individual cannot invade. Since such an individual performs well against the Altruists, stability means that there are sufficiently few Altruists in the population. This result, that conditional kindness (reciprocity) and unconditional kindness (altruism) can blend and co-exist in a stable relationship, when there are not too many of the latter type, is found in many other evolutionary models (see again Sethi and Somanathan, 2003, for a survey).

When part (b) applies, the only advantage of being a Reciprocator rather than a Materialist, namely that one performs well against other Reciprocators, is gone. Indeed, at any population with Altruists the proportion of Materialists is increasing, whereas the opposite is the case for the other two preferences. This implies that in any stable population all individuals defect and are either materialistic or reciprocal.

4 Questions 2 and 3: Comparative-Statics and Welfare Analysis

In this section we assume that $0 < \lambda < 1$, so Proposition 1 holds. Thus the endogenous preference distribution is given by x^*. For simplicity, we set $p = 0$ and $r = 1$. Thus $s < 0$ and $t > 1$. Satisfying (1) and (7) requires $0 < s + t < 1$. Moreover, since, from (8), $\mu = \lambda$, the endogenous preference distribution is

$$x^* = (x_R^*, x_M^*, x_A^*) = \left(\frac{-s(t-1)}{D}, \frac{(1-\lambda)(t-1)}{D}, \frac{-\lambda s}{D} \right), \tag{10}$$

where $D = (\lambda - 1)[1 - (s + t)] - st$.

Whenever the initial preference distribution is not exactly at x^*, the preference distribution fluctuates permanently around x^*. However, it is still very relevant to consider how changes in the exogenous parameters affect the equilibrium preference distribution x^*. This follows because the long-run proportion of preference i along a periodic orbit equals x_i^*. We refer the reader to Hofbauer and Sigmund (1998), Section 7.6.

4.1 Effects of Policies on the Equilibrium Proportion of Altruism (Question 2)

In our simple model we can think of three (government) policies: Those that affect the monetary payoffs, s and t; and those that affect the ability of Reciprocators to co-operate, measured by λ. Since $s < 0$ an increase in s lowers the monetary loss a person suffers when he co-operates and the opponent defects. It can result from a policy that increases society's compensation to individuals who are exploited by opportunistic or criminal behavior. An increase in t raises the monetary returns to a person who defects against an opponent who co-operates. We can think of it as a result of a policy that lowers punishment from opportunistic behavior. Finally, an increase in λ, which increases Reciprocators' ability to co-ordinate on the cooperative equilibrium, can result from policies that increase trust between reciprocal individuals, or from any institutional reform that makes it easier for two reciprocal individuals to cooperate.

Proposition 3 *A policy that, at any given preference distribution,* increases *the Altruist's monetary expected payoff by increasing* s reduces *the equilibrium proportion of altruism:*

$$\frac{\partial x_A^*}{\partial s} < 0.$$

Moreover,

$$\frac{\partial x_A^*}{\partial t} < 0 \ \textit{and} \ \frac{\partial x_A^*}{\partial \lambda} > 0.$$

Proof: See the Appendix.

An increase in the monetary payoff s reduces the Altruist's loss from interacting with a Materialist at the given preference distribution, and hence raises the Altruists' overall expected monetary payoff. Thus one might think that this would increase the equilibrium proportion of Altruists. However, according to the proposition the opposite happens and we may give the following intuitive explanation. When s increases, the Altruist earns more money than the other preference types at the current preference distribution; this triggers an adjustment in the preference distribution. But since the Materialists benefit most from meeting the Altruists, the net result is an increase in *materialism* in society at the expense of the other preferences.

This result shows that when the preference distribution is endogenous, a policy has two effects. One is a 'short run' effect that changes the monetary returns to the various preference types *given* the preference distribution. The other is a 'long run' effect: Due to the changes in the monetary returns obtained by the various preferences there is an adjustment in the preference distribution itself.

It is perhaps more intuitive that an increase in t lowers the equilibrium proportion of altruism. Indeed, an increase in t raises the equilibrium proportions of materialism and reciprocity. The immediate effect is to raise the expected payoff to the Materialists above the payoff of the others. Thus the adjustment dynamic initially raises the proportion of Materialists, at the expense of the two other preference types. However, since the Reciprocator type performs better than the Altruist against the Materialist, the Reciprocator earns more money than the Altruist at the displaced preference distribution. The net effect is that the equilibrium proportion of Reciprocators increases, too.

The last result in the Proposition 3 shows that a policy that increases the ability of reciprocally minded individuals to co-operate promotes altruism: The equilibrium proportion of altruism increases while those of reciprocity and materialism both fall. Intuitively, what happens is that the expected payoff of a Reciprocator has increased at the original equilibrium and this tends to raise the proportion of Reciprocators. However, the Altruists benefit most when meeting a Reciprocator, so the net effect is to increase the equilibrium proportion of Altruists at the expense of the other preference types.

4.2 Altruism and Welfare (Question 3)

If the equilibrium proportion of altruism is affected due to policy changes, how does this affect the individuals' material and subjective equilibrium welfare?

4.2.1 Material Welfare

In the equilibrium preference distribution x^* all preference types have the same expected monetary payoffs. We can therefore use the Materialists' expected money payoff to write the average material equilibrium payoff, denoted π^*, as

$$\pi^* = tx_A^*. \tag{11}$$

Any change in s or λ that affects the equilibrium proportion of altruism influences average material welfare in the same direction (since t is not affected). Thus we ask: Is it possible that, say, an increase in t, even though it lowers the equilibrium proportion of altruism, x_A^*, can raise average material equilibrium welfare, π^*? The answer is no, for we compute

$$\frac{\partial \pi^*}{\partial t} < 0.$$

An increase in t causes such a large decrease in the equilibrium proportion of Altruists, x_A^*, that π^* falls. We therefore have

Proposition 4 *Any policy that increases (reduces) the equilibrium proportion of altruism increases (reduces) average equilibrium material welfare.*

Altruism is always desirable for material prosperity. Nevertheless, as was shown by Proposition 3, some of the policies that accomplish an increase in equilibrium altruism, such as increasing the loss from being defected upon (s), seem quite unattractive in the short run.

4.2.2 Subjective Welfare

Let α_R, α_M, and α_A denote any α satisfying (4), (5), and (6), respectively. Let u_i^*, where $i = R, M, A$, denote the expected subjective payoff to an individual of preference type i when the preference distribution is x^*. Finally, let v denote the subjective payoff to a Reciprocator when interacting with another Reciprocator. Since $v = \lambda$ (this holds since $v = \alpha\lambda + (1-\alpha)\lambda = \lambda$), we have

$$u_R^* = x_R^*\lambda + x_A^* \tag{12}$$

$$u_M^* = x_A^*[\alpha_M t + (1-\alpha_M)s] \tag{13}$$

$$u_A^* = x_R^* + x_M^*[\alpha_A s + (1-\alpha_A)t] + x_A^*. \tag{14}$$

As already mentioned, in the endogenous preference distribution all preference types' expected money earnings are equal to π^*. Moreover, the expected subjective payoff in (12) equals the material expected payoff earned by the Reciprocator type. Thus

$$u_R^* = \pi^*. \tag{15}$$

Moreover, on using (11), we can rewrite (13) and (14) as

$$u_M^* = \pi^* - (1-\alpha_M)(t-s)x_A^* \tag{16}$$

$$u_A^* = \pi^* + (1 - \alpha_A)(t - s)x_M^*. \tag{17}$$

We see from (15) that an increase in Reciprocators' material welfare never lead to lower subjective equilibrium welfare. The reason is that *ex post* distributional matters do not influence the Reciprocators' subjective expected payoff, since in all encounters the Reciprocator gets the same material payoff as his opponent; however, *ex ante* the Reciprocator is clearly motivated by distributional concerns. The second term in each of the expressions (16) and (17) represent the part of subjective equilibrium utility that is influenced by the individual's degree of altruism: The Altruists' and Materialists' expected subjective payoffs are affected by their distributional concerns, via the parameters α_A and α_M. If $\alpha_M < 1$ the Materialist's subjective equilibrium payoff depends negatively on the equilibrium proportion of Altruists; due to his concern for other people's material payoff, the Materialist experiences a utility loss from getting more than other people; it is the effect of remorse or of a 'bad conscience'. Only when the Materialist is completely self-centered, that is, $\alpha_M = 1$, does this concern disappear. If $\alpha_M > 1$, the Materialist actually gets pleasure out of getting more than other people. The Altruist derives pleasure from an increased presence of Materialists, due to the fact that the Altruist gets utility from 'helping' the Materialists to attain a higher monetary payoff.

Let $x^*(a)$ and $x'(a')$ be two preference distributions, each corresponding to a configuration, a and a' respectively, of the exogenous parameters; also, let u_i^* and u_i' denote the subjective welfare of individuals of type i in each preference distribution, respectively. Then, if $u_i^* \leq u_i'$ for all $i = R, M, A$, the change from a to a' gives a Pareto improvement. If all the inequalities are strict, we have a strict Pareto improvement.

A policy change affects Materialists' and Altruists' subjective welfare through two channels: One is through the effect on their personal money earnings and the other is on the induced change in the preference distribution which in turn affects utility, due to the individuals' concern about other people's money earnings. Consider a policy that increases the equilibrium proportion of altruism, and hence (Proposition 4) also raises material equilibrium payoff. The following proposition shows that the positive effect of the policy on subjective welfare (the increase in personal income) can be outweighed by the negative effect on subjective welfare (due to individuals' distributional concerns):

Proposition 5 *A policy of reducing s or t that increases the equilibrium proportion of altruism, x_A^* (and hence increase the average material payoff, π^*) can reduce some individuals' subjective equilibrium welfare if they are 'too' altruistic. More precisely:*

(a). Suppose $\alpha_M > max\{\underline{\alpha}_M, \underline{\alpha}'_M\}$ and $\alpha_A > max\{\underline{\alpha}_A, \underline{\alpha}'_A\}$. Then any policy that raises the equilibrium proportion of altruism gives a strict Pareto-improvement.

(b). If, however, one or more of these inequalities fail to hold, then a policy that increases the equilibrium proportion of altruism reduces some preference types' (Materialist or Altruist) subjective equilibrium welfare.

(c). An increase in λ gives a strict Pareto-improvement.

$$\underline{\alpha}_M = \frac{-s(1 - s - \lambda)}{(1 - s)^2 - \lambda(1 - 2s)}$$

$$\underline{\alpha}'_M = \frac{s\lambda\left[2(t-1)+s\right] - s(2-s)(t-1)}{\lambda(s^2+t-2s-t^2+2st) - 2st + t^2 - s^2 + s^2t + 2s - t}$$

$$\underline{\alpha}_A = \frac{(t-1)\left\{\lambda(t-1+2s) - (t-1)(1-s)\right\}}{\lambda\left\{(1-t)^2 - s(1+s-2t)\right\} - (t-1)^2(1-s)}$$

$$\underline{\alpha}'_A = \frac{\lambda(t-1) + (1-t)^2}{\lambda(2t-1) + (1-t)^2}.$$

Proof: See the Appendix.

It is important to add that these results are fully consistent with our assumption in (3), that individuals always care at least as much for their own as for other people's money earnings. Precisely, there are parameters configurations (s, t, λ) such that $\underline{\alpha}_A > 1/2$ and $\underline{\alpha}'_A M > 1/2$, and such that $\underline{\alpha}_M > (t-1)/(t-s)$ and $\underline{\alpha}'_M > (t-1)/(t-s)$ (cf. (5)).

The result is quite intuitive: As long as the degrees of altruism, $1-\alpha_M$ and $1-\alpha_A$, are not too large, subjective payoffs are predominantly affected by changes in the material payoffs and not by changes in the preference distribution. However, when this is not the case, any policy that affects the money payoff s and t can have adverse effects on subjective welfare.

Consider, for example, a policy that reduces t, the gain from defecting when the other person defects. This increases the material equilibrium payoff, π^*, increases the equilibrium proportion of Altruists, and reduces that of Materialists. When an Altruist is sufficiently altruistic (namely $\alpha_A < \underline{\alpha}_A$), he loses from such a policy. The reason is that her subjective loss from 'helping' the Materialists less (due to the decrease in the equilibrium proportion of materialism in society) outweighs her personal material gain. Moreover, if a Materialist is sufficiently altruistic ($\alpha_M < \underline{\alpha}_M$), he loses, too: His subjective loss from getting more than the Altruists goes up as there are more Altruists and this outweighs his gain from making more money whenever his degree of altruism is sufficiently large. Similarly, a policy that lowers s, and hence increases material prosperity, subjectively hurts the Altruist if he is too caring ($\alpha_A < \underline{\alpha}'_A$): His personal income goes up, but since there are now fewer Materialist he loses more from 'helping' the Materialists less. Similarly, a Materialist can lose: He earns more money due to the policy change, but there are more Altruists in the new preference distribution. Again, if he cares enough about other people's material payoffs ($\alpha_M < \underline{\alpha}'_M$), then his 'bad conscience' from getting more than the Altruists outweighs his personal material gain.

4.3 Comparison with an Exogenous Preference Distribution

Let us now assume that the preference distribution x^* is fixed. It is not difficult to see that (a). An increase in λ, s, or t all give a Pareto improvement and (b). Any exogenous increase in the proportion of Altruists gives a strict Pareto improvement.

These results follow straightforwardly from studying the expressions (12) – (14). As long as changes in money payoffs do not alter the qualitative features of equilibrium behavior in the individual encounters, an increase in the money payoffs always benefit individuals. Part (b) holds because an exogenous increase in the proportion of Altruists always raises the subjective welfare of Reciprocators and Materialists. The same is true for the Altruist, as long as he is not too altruistic, that is $\alpha_A \geq 1/2$; we see from (14) that the Altruist can be hurt from an increase in altruism only of he benefits more from interacting with a Materialist than with an Altruist, that is, if $\alpha_A < (t-1)/(t-s)$. This, however, is not compatible with (3).

We thus see that the results from working with a fixed preference distribution can be very different from those obtained with an endogenous preference distribution: A policy that is beneficial (in the sense of raising material and subjective payoffs) in the short run can be detrimental in the longer run; this is true for any policy that raises s and t. Only a policy that raises λ is unambiguously beneficial in both the short and the long run.

4.4 Normative Implications

The following result is implied by, and summarizes, our previous analysis.

Corollary 1 *Suppose Materialist and Altruist individuals are not too altruistic:* $\alpha_M > max\{\underline{\alpha}_M, \underline{\alpha}'_M\}$ *and* $\alpha_A > max\{\underline{\alpha}_A, \underline{\alpha}'_A\}$. *In order to give a strict Pareto improvement policies should either (i). Increase the monetary loss to an individual from being unilaterally defected on (that is reduce* s*), (ii). Reduce the gain from unilateral defection (reduce* t*), or (iii). Increase the ability of Reciprocators to co-ordinate on the co-operative equilibrium (increase* λ*).*

Again, the perhaps surprising result is part (i): Society can gain from *increasing* the monetary loss from co-operative behavior when the opponent is defecting (make s more negative). Society should not raise compensation to individuals who have been exploited by opportunistic individuals. Intuitively, since an increased compensation would reduce the monetary loss from exploitation, Altruists would outperform the other preference types at the current preference distribution. This would cause a change in the underlying distribution whereby the equilibrium proportion of materialism increases at the expense of altruism and reciprocity (due to the fact that, as explained earlier, the Materialists benefit most from an increase in altruism). Conversely, reducing compensation and so increasing the monetary loss from exploitation would, in the short run, lower the performance of altruism, but would in the longer run 'hurt' materialism most, and raise the equilibrium proportion of altruism. Thus a policy that, in the hope of promoting altruism, raises the material returns from altruism in the short run can actually have the opposite effect in the longer run, due to the effects on the underlying preference distribution.

5 Question 4: Imperfect Information about Preferences

In the previous section we assumed that individuals' preferences were common knowledge. In this section there is imperfect information about preferences: An individual learns only with some probability what the other person's preference type is. Formally, when two players meet, 'Nature' decides whether or not preferences are observable: With probability p, where $0 \leq p \leq 1$, the preferences are common knowledge; with complementary probability, $1-p$, each player observes nothing. We assume that not even the aggregate preference distribution is observed in the latter situation. For a different approach to modeling imperfect information, see for example Engelmann (2001). An interpretation of this set-up is that interaction in society is more or less *anonymous*, the former (latter) situation is modeled by positing a low (high) value of p.

Imperfect type information does not matter for Altruists or Materialists; they always co-operate and defect, respectively. However, when a Reciprocator is not informed about the opponent's preferences, her choice depends on her beliefs about what preference the opponent has, that is, her beliefs about the preference distribution x: We distinguish between two Reciprocator types. Each type has the same preferences, but differ in their beliefs about x. Type 'R_C' has beliefs that make the C action optimal when not observing the opponent's preferences (that is her beliefs about x are such that C is optimal); type 'R_D' plays D in this situation.

We now have four preference types for the evolutionary game, whose matrix is:

	R_C	R_D	M	A
R_C	$p\lambda + 1 - p$	$p\lambda + (1-p)s$	$(1-p)s$	1
R_D	$p\lambda + (1-p)t$	$p\lambda$	0	$p + (1-p)t$
M	$(1-p)t$	0	0	t
A	1	$p + (1-p)s$	s	1

Table 3: the evolutionary game when there is imperfect information about preferences.

From the matrix we immediately obtain

Lemma 1 *Preference type R_C is strictly dominated by preference type R_D for any $p < 1$.*

The interpretation is straightforward: Whenever preferences are not known, the opponent's behavior is independent of one's own preferences. Those who defect in this situation then earn strictly more money than those who co-operate. The lemma implies that we may from the outset ignore preference type R_C. As long as the initial population state is interior, the proportion of R_C individuals approaches zero under the Replicator Dynamic (see Weibull, 1995, Ch. 3 for this result). We can therefore restrict attention to the 3×3 matrix obtained by deleting the first row and first column in the matrix in Table 3.

How does the survival of altruism in the endogenous preference distribution vary with the degree of anonymity, p? When an Altruist meets a Reciprocator, the beneficial strategic effect of the Altruist's preferences on the Reciprocator's behavior is triggered only when the Altruist's preference is recognized; when this does not happen, the Altruist is exploited by the Reciprocator since the latter chooses to defect. It is thus intuitive that altruism only survives when the probability p is sufficiently high. We obtain the following characterization of the endogenous preference distributions.

Proposition 6 *If $p > p'$ then altruism is present in the endogenous preference distribution. If $p < p'$ there is no altruism. More precisely:*

(a). Suppose $p \in [0, p')$. There is no altruism in any stable preference distribution. If exactly $p = 0$, all individuals defect. If $p \in (0, p')$, the all-Reciprocator preference distribution is asymptotically stable (an ESS). Furthermore, if $\lambda = 1$ this is true for all $p \in (0, 1)$.

(b). Let $p \in (p', p'')$ and $\lambda < 1$. There is a unique asymptotically stable preference distribution (an ESS), consisting of a proportion x'_R of Reciprocators and a proportion $1 - x'_R$ of Altruists.

(c). Finally, suppose $p \in (p'', 1]$. There is an asymptotically stable preference distribution, $\tilde{x} = (\tilde{x}_R, \tilde{x}_M, \tilde{x}_A)$, with all three preferences represented.

$$p' = \frac{-s}{1 - s - \lambda} \tag{18}$$

$$p'' = \frac{\lambda - s}{1 - s} \tag{19}$$

$$x'_R = \frac{(1-p)(t-1)}{p[2 - \lambda - (s+t)] - \{1 - (s+t)\}} \tag{20}$$

$$\tilde{x} = (\tilde{x}_R, \tilde{x}_M, \tilde{x}_A) = \left(\frac{-s(t-1)}{E}, \frac{(t-1)[p(1-s) - (\lambda - s)]}{E}, \frac{-s\lambda}{E} \right), \tag{21}$$

where $E = (\lambda - p)[1 - (s+t)] - pst$.

Proof: See the Appendix.

We can interpret part (a) as follows: In a society where people know little, or even nothing, about each other's motivations, altruism cannot survive: The strategic effect of the Altruist's preferences on Reciprocators' behavior is too weak. As long as there is some, but not too much, information ($0 < p < p'$) there is only reciprocity in the population. The reason is that when information is sufficiently low an Altruist performs worse than a Reciprocator against another Reciprocator: The Reciprocator defects, due to the lack of knowledge that he faces an Altruist, and thus ends up involuntarily gaining at the expense of the Altruist. Thus it is an evolutionary advantage for reciprocity not to have too much information about the opponents; for altruism it is a disadvantage not to be recognized. If there is exactly zero information

in society about other people's preferences, Reciprocators cannot distinguish between themselves and Materialists; defection is then the only stable behavior.

Part (b) shows that if there is more information, $p > p'$, then altruism does survive with reciprocity, and there is no materialism. Altruism survives because an Altruist now *is* sufficiently likely to be recognized by a Reciprocator, and hence the strategic effect from which the Altruist benefits is sufficiently strong. Still, howewer, under imperfect information the Reciprocator outperforms an Altruist against another Altruist. These two facts lead to a unique preference distribution at which the two preference types co-exist (we effectively have a 'Hawk-Dove' interaction, with the Altruist being the Dove). Average material payoff in the Reciprocator-Altruist preference distribution x' is

$$\pi' = [p + (1-p)s]x'_R + 1 - x'_R,$$

which is increasing in the population proportion of altruism, $1 - x'_R$.

As information becomes more and more plentiful (p increases), the equilibrium proportion of altruism in the preference distribution increases (x'_R falls). If the probability p exceeds a critical probability, namely p'', there are so many Altruists that a Materialist can invade. Then the qualitative nature of the equilibrium changes and we get asymptotically stable co-existence between all three preference types, as shown in Part (c). The preference distribution $\tilde{x}$ corresponds to x^* for $p = 1$ (cf. Proposition 1): $\lim_{p\to 1^-} \tilde{x} = x^*$. The endogenous preference distribution under perfect preference information can therefore be interpreted as the limiting outcome of the preference distributions under imperfect information when information becomes perfect. However, the stability properties of the equilibrium under perfect information are different from those under imperfect information: Whereas the preference distribution x^* described in Proposition 1 is 'only' stable (and not asymptotically stable), those described in parts (b), (c) and (d) are all asymptotically stable.

5.1 Welfare

We can, as in Section 4, consider how material and subjective welfare varies with the equilibrium proportion of altruism. The average material equilibrium payoff in the preference distribution $\tilde{x}$, given in (21) above, is

$$\tilde{\pi} = t\tilde{x}_A.$$

In the preference distribution x' subjective payoffs can, after some simplifications, be written as

$$u'_R = \pi' - (1-p)(1-x'_R)(1-\alpha_R)(t-s)$$

and

$$u'_A = \pi' + (1-p)x'_R(1-\alpha_A)(t-s).$$

In the preference distribution $\tilde{x}$, subjective payoffs are

$$\tilde{u}_R = \tilde{\pi} - (1-\alpha_R)(1-p)(t-s)\tilde{x}_A$$

$$\tilde{u}_M = \tilde{\pi} - (1-\alpha_M)(t-s)\tilde{x}_A$$

$$\tilde{u}_A = \tilde{\pi} + (1-\alpha_A)(t-s)\left[(1-p)\tilde{x}_R + \tilde{x}_M\right].$$

Once more, as for (15) –(17), in equilibrium subjective welfare goes up when personal material welfare increases, but subjective utility is additionally influenced by other people's earnings. The new thing is that the Reciprocator is affected by the the extent of altruism in society. This is because the Reciprocator sometimes (with probability $1-p$) exploits the Altruist, and this affects the Reciprocator negatively.

In what follows we assume that people are sufficiently self-centered such that subjective equilibrium welfare always move in the same direction as material equilibrium welfare. We therefore consider the effects on material welfare only. Let us first summarize how the average payoff in the endogenous preference distribution depends on p by defining the following function:

$$\pi(p) = \begin{cases} p\lambda & p \in [0, p'] \\ \pi' & p \in (p', p''] \\ \tilde{\pi} & p \in (p'', 1] \end{cases}. \tag{22}$$

Defining π like this at $p = p'$, at $p = p''$, and at $p = 1$ makes sense since $p'\lambda = \lim_{p\to p'+} \pi'$, $\lim_{p\to p''-} \pi' = \lim_{p\to p''+} \tilde{\pi} = \frac{t\lambda}{t-1+\lambda}$ and $\lim_{p\to 1-} \tilde{\pi} = \pi^*$. We then have

Proposition 7 *(a). Consider the preference distribution x', described in Proposition 6. An increase in t reduces material equilibrium welfare, π'; an increase in s increases π', as does an increase in λ and in p.*

(b). Consider the preference distribution $\tilde{x}$. An increase in t reduces material welfare, $\tilde{\pi}$, whereas an increase in s and in λ increases $\tilde{\pi}$. However, an increase in p reduces $\tilde{\pi}$.

(c). Suppose $\lambda < 1$. Then material equilibrium welfare π is strictly increasing at all $[0, p'')$ and strictly decreasing at all $p \in (p'', 1)$. Thus average material equilibrium welfare is maximized at $p = p'' < 1$.

Proof: See the Appendix.

The result that an increase in the sucker's payoff s increases welfare is different from the result for preference distribution x^*, cf. Section 4 (Proposition 3). The reason is the following. In the preference distribution x^* Altruists benefit when s increases, but this is in encounters with Materialists, not with Reciprocators; and Materialists benefit most from meeting the Altruists, whereas the opposite is certainly not true. The net

effect is a reduction in altruism and an increase in materialism. In the preference distribution x' Altruists also benefit from the increase in s, and Reciprocators benefit most from meeting Altruists (like the Materialists do in x^*); however, the difference is that in encounters with Reciprocators it is the Altruist that performs best. Thus the net effect is instead an increase in altruism, which is beneficial for society. An increase in the likelihood that individuals are informed about each other's preferences, p, is also beneficial for society; it magnifies the strategic effect from which Altruists benefit, and increases their presence in the population.

A perhaps surprising result for the preference distribution $\tilde{x}$ is that as individuals obtain more information, average welfare falls; conversely, people would benefit from an *increase* in anonymity. The dynamic effects that give this result are similar to those studied earlier: The immediate short run effect of an increase in information is that Altruists benefit, since they are more often recognized by reciprocal individuals with co-operation as the result. This tends to raise the proportion of Altruists. However, again, since the Materialists benefit most from such an increase the net effect is an increase in materialism, at the expense of altruism and reciprocity.

Part (c) follows from this: When $\lambda < 1$, there is an optimal level of information in society, which equals p'', where $0 < p'' < 1$. Intuitively, there should be sufficiently little information such that the equilibrium proportion of altruism is kept low enough that Materialists cannot enter the population; or, if the Materialists are already present ($p > p''$), their frequency should be driven to zero (p approaches p'' from above). As long as the degree of information is below the critical level p'', society benefits from making more information available, since this increases the equilibrium proportion of altruism, which is beneficial. These results show that in addition to affecting the money payoffs, s and t, and the ability of reciprocal individuals to co-operate, measured by λ, society can also influence the equilibrium proportion of altruism by affecting the degree of anonymity in society, measured by p.

6 Related Evolutionary Literature

Altruism is studied in several other evolutionary models, such as Bester and Güth (1998) [their results are extended in Bolle (2000) and Possajennikov (2000)], Bergstrom and Stark (1993), Bergstrom (1995), Berninghaus, Korth, and Napel (2007), and Eshel, Samuelson, and Shaked (1998). See also the excellent survey in Bergstrom (2002). Poulsen and Poulsen (2006) analyze the evolution of preferences in a 'Game of Life' where individuals are sometimes involved in simultaneous and at other times in sequential interaction; that paper contains a result qualitatively similar to Proposition 1, but contains no welfare analysis. Other evolutionary models of preference evolution are Engelmann (2001), Fehrstman and Weiss (1998), and Ockenfels (1993).

Bester and Güth assume that information about preferences is perfect and show that when actions are strategic complements, a population where all individuals display the same positive degree of altruism is evolutionarily stable. If, on the other hand, actions are strategic substitutes, then only selfishness survives. This is because these relationships determine the strategic effect (Schelling, 1978) of the altruist's preferences on other individuals' behavior, and this is crucial for evolutionary sta-

bility. There are several differences between their model and our model. First, in their model there is always a unique Nash equilibrium when any two individuals meet. In our model there are multiple equilibria when reciprocal individuals meet, and we study how the endogenous preference distribution depends, among other things, on how this selection problem is overcome. Second, in Bester and Güth's model all individuals display the same degree of altruism in the evolutionarily stable outcome. Our main result, on the other hand, is that there can be preference heterogeneity, where individuals with different degrees of altruism live side by side in society.

Third, we employ a more restrictive notion of altruism than Bester and Güth: In their set-up altruism is defined in terms of people's preferences, not in terms of behavior: Any individual whose utility function assigns any weight to other people's money earnings is labeled 'altruist', no matter his actual behavior. Thus our Materialist preference type is an 'Altruist' according to their model, as long as $\alpha_M < 1$. In our set-up, on the other hand, altruism is defined in terms of actual behavior: An 'Altruist' is a person who *acts* altruistically (unconditionally co-operates).

7 Summary

We analyze the survival ability and welfare effects of altruism in a model where preferences are endogenous. Altruism can indeed materially 'pay off' and thus survive. We show that the endogenous preference distribution can contain both reciprocal, materialistic and altruistic individuals. We analyze how changes in the money payoffs affect this preference distribution. When individuals are not 'too' altruistic an increase in the equilibrium proportion of altruism in society is always desirable, in the sense that is raises society's material and subjective welfare. If, on the other hand, individuals care a lot about other individuals' money earnings, policies that increase material welfare can reduce subjective welfare. Another insight is that a policy that increases the monetary returns to altruism, which on its own would increase material and subjective welfare, can give more materialism in society in the longer run, which reduces welfare. These, and other, of our welfare results, show that the conclusions from an analysis with an endogenous preference distribution can be very different from those based on a (short-run) assumption of fixed preferences. We also show that our results continue to hold when preferences are only imperfectly observable.

8 Appendix

Proof of Proposition 1: At $x = (x_R, x_M, x_A)$ the expected money payoffs to the R, M, and A types are: $\pi(R, x) = \mu x_R + r x_A + p x_M$, $\pi(M, x) = p x_R + p x_M + t x_A$ and $\pi(A, x) = r x_R + s x_M + r x_A$. Solving the system $\pi(R, x) = \pi(M, x)$ and $\pi(M, x) = \pi(A, x)$, using $x_A = 1 - x_R - x_M$, gives the solution in (9). It is straightforward to verify that $0 < x_i^* < 1$ for all $i = R, M, A$ under the stated conditions on the money payoffs and μ. There are four equilibria for the Replicator dynamic: The three vertices $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, and x^*. We examine the stability of each in turn. The vertices $(1, 0, 0)$ and $(0, 0, 1)$ are not Nash equilibria, hence they are unstable. The equilibrium $(0, 1, 0)$ is, in fact, a Nash equilibrium, but is unstable. To see this suppose

that some small perturbation takes $(0, 1, 0)$ to $x' = (\epsilon, 1-\epsilon, 0)$ where $\epsilon > 0$ is a small number. At any such x', the R-type earns a strictly higher expected monetary payoff than M. Thus at x' the proportion of R types increases at the expense of M types, which is a contradiction of stability. Turning attention to the interior equilibrium x^*, we may first, instead of the matrix in the main text, study the equivalent matrix, obtained by deleting the number μ (p) $[r]$ from all entries in the first (second) [third] column:

	R	M	A
R	0	0	0
M	$p-\mu$	0	$t-r$
A	$r-\mu$	$s-p$	0

Or, in abbreviated form,

	R	M	A
R	0	0	0
M	α	β	γ
A	δ	ϵ	θ

Denote this matrix by A, with typical element a_{ij}, with $i, j = R, M, A$, and let $x = (x_R, x_M, x_A)$ be a column vector. Hofbauer (1981) shows that the three-dimensional Replicator dynamic, $\dot{x} = x\left[Ax - x^T Ax\right]$, is equivalent to the two-dimensional Lotka-Volterra dynamic:

$$\dot{x} = x\left[a_{MR} + a_{MM}x + a_{MA}y\right]$$

$$\dot{y} = y\left[a_{AR} + a_{AM}x + a_{AA}y\right],$$

where $x = x_M/x_R$ and $y = x_A/x_R$. If (x, y) is an interior equilibrium of the Lotka-Volterra dynamic, then

$$x^* = (x_R^*, x_M^*, x_A^*) = (1/(1+x+y), x/(1+x+y), y/(1+x+y)) \tag{23}$$

is an equilibrium for the Replicator Dynamic. Conversely, if x^* is an interior equilibrium for the Replicator Dynamic, then

$$(x, y) = (x_M^*/x_R^*, x_A^*/x_R^*) \tag{24}$$

is an equilibrium for the Lotka-Volterra dynamic. Moreover, results about the stability of equilibria for the Lotka-Volterra dynamic carry over to the Replicator dynamic and conversely, via the two transformations given above. We refer the reader to Hofbauer and Sigmund (1998), Section 7.5, for details. Our strategy is to use the transformation (24) and study the stability of the equilibrium (x, y) for the Lotka-Volterra dynamic; the equilibrium x^* then has the same stability properties under the Replicator Dynamic that (x, y) has for the Lotka-Volterra dynamic.

On using the transformation given above on x^*, we compute the following equilibrium for the Lotka-Volterra dynamic:

$$x = \frac{(t-r)(r-\mu)}{(t-r)(p-s)} = \frac{r-\mu}{p-s} \quad \text{and} \quad y = \frac{(p-\mu)(s-p)}{(t-r)(p-s)} = \frac{\mu-p}{t-r}.$$

We may verify that when these expressions are used in (23), we indeed obtain the solutions in (9).

In characterizing the stability of (x, y), we may use the very useful results given in Bomze (1983). Bomze shows that if the three expressions $\beta\theta - \gamma\epsilon$, $\alpha\epsilon - \beta\delta$ and $\gamma\delta - \alpha\theta$ have the same sign, and if $\beta x + \theta y = 0$, then (x, y) is a center for the Lotka-Volterra system. We have $\beta\theta - \gamma\epsilon = 0 - (t-r)(s-p) > 0$, $\alpha\epsilon - \beta\delta = (p-\mu)(s-p) - 0 > 0$ and $\gamma\delta - \alpha\theta = (t-r)(r-\mu) - 0 > 0$, using the fact that $p < \mu < r$. Moreover, since $\beta = \theta = 0$, it follows that $\beta x + \theta y = 0$. We may therefore conclude that (x, y) is a center for the Lotka-Volterra system. Using the results in Hofbauer (1981) mentioned above allows us to conclude that our equilibrium x^* is a center. ■

Proof of Proposition 2: (a). Let x denote any population where individuals are Reciprocators (R) and/or Altruists (A). The average payoff in any such population is $\pi(x, x) = r$, since all individuals co-operate with each other. Any population x', where a proportion x'_M are M types, earn expected payoff $\pi(x', x) = (1 - x'_M)r + x'_M[(1 - x_A)p + x_A t]$ against x. Thus when $(1 - x_A)p + x_A t \leq r$, that is $x_A \leq (r-p)/(t-p)$, we have $\pi(x', x) \leq \pi(x, x)$ for all populations x', so x is a Nash equilibrium. To show that x is a Neutrally Stable Strategy (Maynard-Smith, 1982) we must show that $\pi(x, x') \geq \pi(x', x')$ for any x' satisfying $\pi(x', x) = \pi(x, x)$. However, since for any such x' we have $\pi(x, x') = \pi(x', x') = r$, this is satisfied. (b). Let x denote any population with a strictly positive proportion of Altruists. To show that no such x is an NSS, it suffices to show that x is not a Nash equilibrium. First, if x uses the A preference only, then M is clearly a better reply. Second, if x is a mix of preference A and R, then A is a better reply. Finally, if x uses all three strategies, then M earns strictly higher expected payoff than R. Thus we may conclude that there is no Nash equilibrium where preference A is used. ■

Proof of Proposition 3: From (10) we compute the following partial derivatives:

$$\frac{\partial x^*_A}{\partial s} = \frac{-\lambda(1-\lambda)(t-1)}{D^2} < 0,$$

and

$$\frac{\partial x^*_A}{\partial t} = \frac{s\lambda(1-s-\lambda)}{D^2} < 0,$$

since $s < 0$ and $\lambda < 1 - s$. Moreover,

$$\frac{\partial x^*_A}{\partial \lambda} = \frac{-s(t-1)(1-s)}{D^2} > 0,$$

since $(t-1)(1-s) > 0$. ■

Proof of Proposition 5: Effects of Changes in t: From (15) we obtain $\partial u^*_R/\partial t < 0$. From (13),

$$\frac{\partial u_M^*}{\partial t} = \frac{\alpha_M s\lambda(1-s)(1-\lambda))}{D^2} + \frac{(1-\alpha_M)s^2\lambda(1-s-\lambda)}{D^2}.$$

Upon simplifying, this can be written as

$$\frac{s\lambda}{D^2}\left[\alpha_M\left\{1-2s+s^2+2s\lambda-\lambda\right\}+s(1-s-\lambda)\right].$$

Thus $\partial u_M^*/\partial t < 0$ if $\alpha_M\left\{1-2s+s^2+2s\lambda-\lambda\right\} > -s(1-s-\lambda)$. The right hand side of this inequality is strictly positive. It is thus necessary that $1-2s+s^2+2s\lambda-\lambda > -s(1-s-\lambda)$. This is the same as $1-s > \lambda(1-s)$, which holds. Thus $\partial u_M^*/\partial t < 0$ iff

$$\alpha_M > \frac{-s(1-s-\lambda)}{(1-s)^2-\lambda(1-2s)} \equiv \underline{\alpha}_M.$$

We verify that $\underline{\alpha}_M < 1$ is equivalent to $\lambda(1-s) < 1-s$. This always holds.

Next, on inserting (11) and the expression for x_A^* into (17), differentiating and re-arranging, we obtain

$$\frac{\partial u_A^*}{\partial t} = \frac{(1-\lambda)}{D^2}\left[\alpha_A A - B\right],$$

where $A = \lambda\left\{1-s^2+t^2+2st-2t-s\right\}+2t+st^2-1-t^2+s-2st$ and $B = (t-1)\left\{\lambda(t-1+2s)-(t-1)(1-s)\right\}$. Thus $\partial u_A^*/\partial t < 0$ iff

$$\alpha_A A < B. \tag{25}$$

We first consider the sign of B. $B < 0$ iff $\lambda(t-1+2s) < (t-1)(1-s)$. It is sufficient for this that $2s < 1-t$, which, in turn, is implied by the fact that $s+t < 1$. Thus $B < 0$. It is thus necessary for $\partial u_A^*/\partial t < 0$ that $A < 0$. $A < 0$ iff

$$\lambda\left\{1+t^2+2st-2t-s-s^2\right\} < 1+t^2+2st-2tst^2.$$

The right hand side is strictly positive (it can be factored to $(t-1)^2(1-s)$), so a sufficient condition for $A < 0$ is $s^2 < st^2$, which holds. Thus $A < 0$. A necessary condition for (25) is therefore $A < B$. This is equivalent to $\lambda s(s-1) > 0$. Since $s(s-1) > 0$, this holds. It then follows from the signs of A and B that $\partial u_A^*/\partial t < 0$ iff

$$\alpha_A > \frac{B}{A} \equiv \underline{\alpha}_A. \tag{26}$$

It is necessary for (26) that (cf. (6)) $\underline{\alpha}_A < t/(t-s)$ or $(t-s)B < tA$. Writing this out and collecting terms in λ gives

$$\lambda\left\{1-t+t^2+st-2s\right\} < (t-1)^2(1-s).$$

Since the right hand side is strictly positive, this holds for sufficiently small $\lambda > 0$.

Verifying when the conditions $\alpha_M > \underline{\alpha}_M$ and $\alpha_A > \underline{\alpha}_A$ bind: First, given fixed values of α_A, α_R and α_M, with $1/2 \leq \alpha_A < \alpha_R < \alpha_M$, s and t must satisfy the constraint $0 < s + t < 1$, together with the constraints (4) – (6).

Analyzing these constraints reveals that, given t, s must satisfy

$$max\left\{-\frac{1-\alpha_A}{\alpha_A}t, \frac{1}{1-\alpha_M}[1-\alpha_M t]\right\} < s < min\left\{-\frac{1-\alpha_R}{\alpha_R}t, \frac{1}{1-\alpha_R}[1-\alpha_R t]\right\}, \tag{27}$$

for $t \in (\underline{t}, \bar{t})$, where

$$\underline{t} = \frac{\alpha_R}{\alpha_M + \alpha_R - 1}$$

and

$$\bar{t} = \frac{\alpha_A}{\alpha_R + \alpha_R - 1}.$$

The set of s values that satisfy (27) is non-empty.

Effects of Changes in s:

We compute

$$\frac{\partial u_m^*}{\partial s} = \alpha_M t \frac{\partial x_A^*}{\partial s} + (1-\alpha_M)\frac{\partial}{\partial s}\left\{s x_A^*\right\}, \tag{28}$$

where $\partial x_A^*/\partial s$ is given above and

$$\frac{\partial}{\partial s}\left\{s x_A^*\right\} = \frac{-\lambda s\left[-\lambda\left\{2(t-1)+s\right\} + (t-1)(2-s)\right]}{D^2}.$$

Re-arranging (28) gives

$$\frac{\partial u_M^*}{\partial s} = \frac{-\lambda}{D^2}\left[\alpha_M A - B\right],$$

where

$$A = \lambda\left\{2st - 2s + s^2 - t^2 + t\right\} + t^2 - t - 2st + 2s + s^2 t - s^2, \tag{29}$$

and $B = s\left\{\lambda[2(t-1)+s] - (t-1)(2-s)\right\}$. Thus $\partial u_M^*/\partial s < 0$ iff $\alpha_M A > B$. From (29) $A > 0$ iff

$$\lambda\left\{s^2 + t - 2s - t^2 + 2st\right\} > 2st - t^2 + s^2 - s^2 t - 2s + t. \tag{30}$$

The right hand side can be factored to $(t-1)[2s-s^2-t]$ and is thus negative iff $s^2-2s+t>0$. Since $s<0$, this always holds. Thus the right hand side in (30) is strictly negative. We next note that the expression in the bracket on the left hand side is strictly larger than the expression on the right hand side. Thus $A>0$. $B<0$ whenever

$$\lambda\left\{2(t-1)+s\right\}>(2-s)(t-1).$$

Since the right hand side is strictly positive, it is necessary for this inequality that $2(t-1)+s>(2-s)(t-1)$ or simply $s(t-1)>-s$. Since $s<0$, this is not feasible. Thus $B>0$ and so $\partial u_M^*/\partial s<0$ iff $\alpha_M>B/A\equiv\underline{\alpha}_M$.

Next, from (17) a sufficient condition for $\partial u_A^*/\partial s<0$ is $\frac{\partial}{\partial s}\left\{(t-s)x_M^*\right\}<0$, where

$$\frac{\partial}{\partial s}\left\{(t-s)x_M^*\right\}=\frac{-(\lambda-1)(t-1)}{D^2}\left[\lambda(2t-1)+(1-t)^2\right].$$

This, however, is strictly negative, so we need to study the entire expression in (17). We obtain

$$\frac{\partial u_A^*}{\partial s}=t\frac{\partial x_A^*}{\partial s}+(1-\alpha_A)\frac{\partial}{\partial s}\left\{(t-s)x_M^*\right\}.$$

Inserting the expressions for the derivatives and simplifying yields $\frac{\partial u_A^*}{\partial s}=\frac{(1-\lambda)(t-1)}{D^2}\left[B-\alpha_A A\right]$, where $A=\lambda(2t-1)+(1-t)^2$ and $B=\lambda(t-1)+(1-t)^2$. Since $A>0$ and $B>0$, $\frac{\partial u_A^*}{\partial s}<0$ iff $\alpha_A>\frac{B}{A}\equiv\underline{\alpha}'_A$.

Verifying when the conditions $\alpha_M>\underline{\alpha}'_M$ and $\alpha_A>\underline{\alpha}'_A$ bind: Considering the Materialist type, a necessary condition for $\partial u_M^*/\partial s>0$ is

$$\frac{1-s}{t-s}<\underline{\alpha}_M,$$

that is,

$$(1-s)A<(t-s)B. \tag{31}$$

A sufficient condition for this would be $B>A$. This is equivalent to $\lambda t(t-1)>t(t-1)$, which is not feasible. Thus $0<B<A$. Next, on isolating λ in (31) we obtain

$$\lambda\left\{-3st+s^2t+st^2-s^2-t+2s+t^2\right\}>st^2-3st-s^2t^2+2s^2t-s^2+t^2+2s-t. \tag{32}$$

We observe that the expression in the bracket on the left hand side is strictly larger than the expression on the right hand side, iff $s^2t>2s^2t-s^2t^2$. This is the same as $t>1$, which always holds. Thus we may conclude that whenever the expression on

the right hand side is strictly positive, then (32) holds for a sufficiently large λ. And if the right hand side is negative, then (32) holds for all $\lambda \in (0,1)$.

For $\partial u_A^*/\partial s > 0$ it is necessary that $\underline{\alpha}'_A > 1/2$, that is, $2B > A$. Inserting the expression for A and B and simplifying gives

$$\lambda < (1-t)^2, \tag{33}$$

which, given t, holds for a sufficiently small λ. Thus, when (33) holds, $\partial u_A^*/\partial s > 0$ for $\alpha_A \in (1/2, \underline{\alpha}'_A)$.

Effects of Changes in λ: From (15) and the fact that $\partial \pi^*/\partial \lambda > 0$ we again have $\partial u_R^*/\partial \lambda > 0$. Furthermore, we may, from (13), conclude that $\partial u_M^*/\partial \lambda > 0$. Next, we see from (17) that the effect of λ on u_A^* is ambiguous: An increase in λ gives an increase in π^*. However, the increase in λ also gives a decrease in x_M^*. Inserting the expressions for x_A^* and x_M^* into (17) and re-arranging gives

$$u_A^* = \frac{-\lambda[(1-\alpha_A)(t-s)(t-1)+st] + (1-\alpha_A)(t-s)(t-1)}{D}.$$

Upon differentiating and simplifying, we obtain

$$\frac{\partial u_A^*}{\partial \lambda} = \frac{\{B+st\}\left[\lambda(1-s-t)-D\right] - B(1-s-t)}{D^2},$$

where $B = (1-\lambda)(t-s)(t-1)$. This can be simplified to

$$\frac{\partial u_A^*}{\partial \lambda} = \frac{-st(t-1)\left[\alpha_A(t-s)-(t-1)\right]}{D^2}.$$

Since $-st(t-1) > 0$, we may conclude that $\partial u_A^*/\partial \lambda > 0$ iff $\alpha_A > (t-1)/(t-s)$. Since $(t-1)/(t-s) < 1/2$, due to (1), (3) ensures that this holds. ■

Proof of Proposition 6: (a). Suppose first $p = 0$. There can be no Altruists in any stable preference distribution, for the Altruists always co-operate and the Materialists always defect, so in any population with Altruists the proportion of Materialists would increase, contradicting stability. This then implies that in any stable preference distribution all players defect, that is they are Materialists and/or Reciprocators. Moreover, sufficiently many must be Materialists, since otherwise the Reciprocators would choose to co-operate, contradicting stability (full details are available from the authors upon request). Suppose then $p \in (0, p')$. We verify that the R preference type is a strict Nash equilibrium when $p\lambda > p + (1-p)s$ or $p > p'$, where p' is given in (18). Then the R preference is an ESS and hence asymptotically stable for the Replicator Dynamic.

(b). Let x denote the preference distribution where a proportion x_R are R types and the remaining proportion $1 - x_R$ are A types. Expected payoffs to R and A are $p\lambda x_R + [p + (1-p)a](1-x_R)$ and $[p + (1-p)s]x_R + 1 - x_R$. The expected payoffs are equal exactly at $x_R = x'_R$, given in (20). To ensure $x'_R < 1$ it is necessary that $p > p'$,

given earlier. Let $(x'_R, 0, 1-x'_R) \equiv x'$. Then $\pi(x', x') = [p + (1-p)b]x'_R + 1 - x'_R$. To show that x' is a Nash equilibrium, we consider the expected payoff to the M preference type, which is $\pi(M, x') = t(1 - x'_R)$. Some straightforward but tedious calculations show that $\pi(x', x') - \pi(M, x')$ can be written as

$$\pi(x', x') - \pi(M, x') = \frac{p(t-1)[\lambda - s - p(1-s)]}{p[2 - \lambda - (s+t)] - \{1 - (s+t)\}}.$$

Since the denominator is strictly positive, this is positive when $p \leq p''$, where p'' is given in (19). Then (x', x') is a Nash equilibrium. We also verify that $p' < p''$ is equivalent to $0 < \lambda(1-\lambda)$, which holds for all $\lambda \in (0,1)$. If $\lambda = 1$, (x', x') is not a Nash equilibrium. We next show that x' is an ESS whenever $p \in (p', p'')$. Since $p < p''$ we have $\pi(x, x') < \pi(x', x')$ for any x that has M in its support. Since $\pi(x, x') = \pi(x', x')$ for any x with only R or A in the support, to show that x' is an ESS we need to verify that $\pi(x', x) > \pi(x, x)$ for all such x. We can write the difference $\pi(x', x) - \pi(x, x)$ as

$$(x'_R - x_R)\left[p\lambda x_R + [p + (1-p)t](1 - x_R)\right] + (1 - x'_R - (1 - x_R))\left[[p + (1-p)s] + 1 - x_R\right]$$

or as

$$(x'_R - x_R)\left[-x_R\left\{p[2 - \lambda - (s+t)] - [1 - (s+t)]\right\} + (1-p)t - (1-p)\right].$$

Using (20), we see that the expression in the bracket can be written as $-x_R + x'_R$. Thus

$$\pi(x', x) - \pi(x, x) = (x'_R - x_R)^2.$$

This is strictly positive for any $x_R \neq x'_R$. Thus x' is an ESS.

(c). For a completely mixed Nash equilibrium to exist it is necessary that A is a best reply to R, i.e., $p > p'$. On solving the equations giving the expected payoffs, we obtain the solution $\tilde{x}$ in (21). We next investigate when $0 < \tilde{x}_i < 1$ for $i = R, M, A$. To verify that $0 < \tilde{x}_R$ we must show that the denominator is strictly positive, that is, $p[1 - s - t + st] < \lambda[1 - s - t]$; however, since $1 - s - t + st < 0$, this holds. Next, $\tilde{x}_R < 1$ is equivalent to $p[1-s-t+st] < \lambda[1-(s+t)] - s(1-t)$. A sufficient condition is that it holds at $p = 1$, namely $1 - t < \lambda(1 - s - t)$. Since $1 - s - t > 0$, this holds. Next, considering $\tilde{x}_M$ we see that the numerator is strictly positive, and so $\tilde{x}_M > 0$, when $p > p''$. $\tilde{x}_M < 1$ iff $0 < \lambda[1 - (s+t)] + (t-1)(\lambda - s)$. Since $1 - (s+t) > 0$, this holds for all $\lambda > 0$. Finally, $\tilde{x}_A < 1$ is the same as $0 < \lambda(1-t) - p(1 - s - t) - spt$; simplifying further gives $p(t-1)(1-s) > \lambda(t-1)$, or $p > \lambda/(1-s)$. A sufficient condition for this is $\lambda/(1-s) < p''$ or $\lambda < \lambda - s$, which clearly holds. Thus $\tilde{x}_A < 1$.

To show that $\tilde{x}$ is asymptotically stable, we proceed as in the proof of Proposition 1. The entries in the equivalent matrix are: $\alpha = -p\lambda$, $\beta = 0$, $\gamma = (t-1)p$, $\delta = p + (1-p)s - p\lambda$, $\epsilon = s$ and $\theta = 1 - p - (1-p)t = (1-p)(1-t)$. Using the results in Bomze (1983), we need to consider the signs of $\beta\theta - \gamma\epsilon = -s(t-1)$, $\alpha\epsilon - \beta\delta = -sp\lambda$ and $\gamma\delta - \alpha\theta = p(t-1)[p(1-s) - (\lambda - s)]$. The first two expressions are strictly

positive. The third is strictly positive when $p > p''$, which holds by assumption. Next, since $x = [\lambda - s - p(1-s))]/s$ and $y = \lambda/(t-1)$, we get $\beta x + \theta y = \lambda(p-1)$. This is strictly negative for all $\lambda > 0$ and $p < 1$. This implies that $\tilde{x}$ is asymptotically stable for the Replicator Dynamic. ■

Proof of Proposition 7: (a). It can be seen directly from the expression for x'_R that an increase in t reduces x'_R (the numerator is decreasing in t and the denominator is increasing) and hence raises the equilibrium proportion $1 - x'_R$ of altruism. Since equilibrium material average welfare, $[p+(1-p)s]x'_R + 1 - x'_R$, is decreasing in x'_R, the result follows. With respect to changes in s (λ), it is again quite straightforward to see that x'_R is decreasing in s (λ), so the same conclusion as for changes in t follows. Regarding the effects of changes in p, we obtain

$$\frac{\partial x'_R}{\partial p} = \frac{p-1}{E} - \frac{(-1+p)^2(-1+t)}{E^2}.$$

This can be simplified to

$$\frac{\partial x'_R}{\partial p} = \frac{(-1+p)(-p-s+sp+p\lambda)}{E^2}.$$

This derivative is strictly positive iff $p > p'$, which holds by assumption.

(b). Changes in t: With $\tilde{\pi} = t\tilde{x}_A$, we compute

$$\frac{\partial \tilde{\pi}}{\partial t} = \frac{-s\lambda}{E} - \frac{ts\lambda\,(-p+sp+\lambda)}{E^2}$$

or, after some simplifications,

$$\frac{\partial \tilde{\pi}}{\partial t} = -\frac{s\lambda\,(sp+\lambda-s\lambda-p)}{E^2}.$$

This expression is strictly negative iff $sp + \lambda - s\lambda - p < 0$ or $p > \lambda$. A sufficient condition for this is $\lambda < p''$; this can be simplified to $-\lambda s < -s$ or $\lambda < 1$. Thus $\partial\tilde{\pi}/\partial t < 0$ for all $\lambda < 1$. Considering the effects of s, we compute

$$\frac{\partial \tilde{\pi}}{\partial s} = -\frac{t\lambda\,(tp-t\lambda+\lambda-p)}{E^2},$$

which is strictly negative if $tp - t\lambda + \lambda - p > 0$ or simply $p > \lambda$; the argument just given above shows that this holds. With respect to λ, we obtain

$$\frac{\partial \tilde{\pi}}{\partial \lambda} = \frac{tsp\,(1-s-t+st)}{E^2}.$$

This is strictly positive, since $1 - s - t + st < 0$ is the same as $(1-t)(1-s) < 0$, which clearly holds. Consider finally the effect of changes in p: The denominator in the expression for $\tilde{x}_A$ can be written as $E = \lambda[1-(s+t)] - p[1-(s+t)+st]$. Since $1-(s+t)+st < 0$ is equivalent to $(1-t)(1-s) < 0$, which clearly holds, we have $\partial\tilde{x}_A/\partial p < 0$. Since $\tilde{\pi} = t\tilde{x}_A$, the conclusion follows. Part (c). This follows from parts (a) and (b). ■

9 References

Andreoni, J. (1990): "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving", Economic Journal, 100, 464-477.

Becker, G. (1974): "A Theory of Social Interactions", Journal of Political Economy, 82, 1063-1093.

Becker, G. (1976): "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology", Journal of Economic Literature, 14 (3), 817-26.

Bergstrom, T. (1989): "A Fresh Look at the Rotten Kid Theorem–and Other Household Mysteries", Journal of Political Economy, 97 (5), 1138-59.

Bergstrom, T. (1995): "On the evolution of altruistic ethical rules for siblings", American Economic Review, 85, 58-81.

Bergstrom, T. (2002): "Evolution of Social Behavior: Individual and Group Selection", Journal of Economic Perspectives, 16 (2), 67-88.

Bergstrom, T. and Stark, O. (1993): "How altruism can prevail in an evolutionary environment", American Economic Review (Papers and Proceedings), 83, 149-155.

Bernheim, D.: and Stark, O. (1988): "Altruism Within the Family Reconsidered: Do Nice Guys Finish Last?", American Economic Review, 78 (5), 1034-1045.

Bernheim, D., Shleifer, A. and Summers, L. (1985): "The Manipulative Bequest Motive", Journal of Political Economy, 93, 1045-1076.

Berninghaus, S., Korth, C., and Napel, S. (2007): "Reciprocity – an indirect evolutionary analysis", Journal of Evolutionary Economics, 17, 579-603.

Bester, H. and Güth, W. (1998): "Is Altruism Evolutionarily Stable?", Journal of Economic Behavior and Organization, 34, 193-209.

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

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

Bruce, N. and Waldman, M. (1990): "The rotten-kid theorem meets the Samaritan's dilemma", Quarterly Journal of Economics, 105, 155-165.

Buchanan, J.M. (1972): The Samaritan's Dilemma, reprinted in: Buchanan, J.M. (1977): Freedom in Constitutional Contract, Texas A& M University Press: 169-185.

Collard, D. (1978): Altruism and Economy, Martin Robertson and Co. Ltd.

Edgeworth, F. (1967): Mathematical Physics (1881), Reprints of Economic Classics, Augustus M. Kelley Publishers.

Engelmann, D. (2001): "Asymmetric Type Recognition with Applications to Dilemma Games", Metroeconomica, 52(4), 357-375.

Eshel, I., Samuelson, L. and Shaked, A. (1998): "Altruists, Egoists and Hooligans

in a Local Interaction Model", American Economic Review, 88, 157-179.

Fehr, E. and Fischbacher, U. (2002): "Why social preferences matter - the impact of non-selfish motives on competition, cooperation and incentives", Economic Journal, 112, C1-C33.

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.

Gambetta, D. (1988): "Can we Trust Trust?", in Gambetta, D. (ed): Trust: Making and breaking Cooperative Relations, Basil Blackwell.

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. (1998): "The indirect evolutionary approach: Bridging the gap between rationality and adaptation", Rationality and Society, 10, 377-399.

Güth, W. and Napel, S: (2006): "Inequality Aversion in a Variety of Games - An Indirect Evolutionary Analysis", Economic Journal, 116, 1037-1056.

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

Hirshleifer, J. (1977): "Shakespeare vs. Becker on Altruism: The Importance of Having the Last Word", Journal of Economic Literature, 15, 500-502

Hochman, H.M. and Rodgers, J.D. (1969): "Pareto optimal redistribution", American Economic Review, 59, 542-557.

Hofbauer, J. and Sigmund, K. (1998): Evolutionary Games and Population Dynamics, Cambridge University Press.

Königstein, M. and Müller, W. (2000): "Combining Rational Choice and Evolutionary Dynamics: The Indirect Evolutionary Approach", Metroeconomica, 51 (3), 235-256.

Maynard-Smith, J. (1982): Evolution and the Theory of Games, Cambridge University Press.

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

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

Poulsen, A. and Poulsen, O. (2006): "Endogenous Preferences and Social Dilemma Institutions", Journal of Institutional and Theoretical Economics, 162 (4), 627-660.

Rotemberg, J. J. (1994): "Human Relations in the Workplace", Journal of Political Economy, 102 (4), 684-717.

Schelling, T. (1978): "Altruism, Meanness, and Other Potentially Strategic Behaviors", American Economic Review (Papers and Proceedings), 68, 229-230.

Sen, A. (1967): "Isolation, Assurance and the Social rate of Discount", Quarterly Journal of Economics, 81, 112-125.

Sethi, R. and Somanathan, E. (2003): "Understanding reciprocity", Journal of Economic Behavior and Organization, 50, 1-27.

Stark, O. (1989): "Altruism and the Quality of Life", American Economic Review, 79 (2), 86-90.

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

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

Witt, U. (1991): "Economics, Sociobiology, and Behavioral Psychology on Preferences", Journal of Economic Psychology, 12, 557-573.