Preference Evolution and Reciprocity*

Rajiv Sethi[†] E. Somanathan[‡]

March 12, 1999

Abstract

This paper provides an evolutionary theory of reciprocity as an aspect of preference interdependence. It is shown that *reciprocal preferences*, which place negative weight on the payoffs of materialists and positive weight on the payoffs of sufficiently altruistic individuals can invade a population of materialists in a class of aggregative games under both individual selection and random matching. Such preferences are efficiency-reducing when they are rare and efficiency-enhancing when they are widespread, suggesting that they can persist even under group selection and assortative matching. In comparison with simpler specifications of preference interdependence (such as pure altruism or envy), the survival of such preferences is therefore less sensitive to details of the evolutionary selection process.

JEL Classification: C72, D62.

Keywords: Reciprocity, Evolution, Preference Interdependence.

*We thank Levent Koçkesen and Efe Ok for extensive discussions on the subject matter of this paper, participants at the University of Massachusetts conference on *Modeling Strong Reciprocity* for comments on an earlier version, and the National Science Foundation for financial support.

[†]Department of Economics, Barnard College, Columbia University, 3009 Broadway, New York, NY 10027 (rs328@columbia.edu).

[‡]Department of Economics, University of Michigan, 611 Tappan Street, Ann Arbor, MI 48109-1220 (esomana@umich.edu).

1 Introduction

Experimental support for the standard conception of the economic actor as a creature driven by material self-interest has, at best, been mixed. Predictions made on the basis of this conception accord closely with the behavior of subjects in some environments, such as competitive auctions and market games (Smith, 1982, Roth *et al.*, 1991), but fail rather dramatically in others, such as public goods, ultimatum bargaining, and gift exchange games (Isaac and Walker, 1988, Güth *et al.*, 1982, Fehr *et al.*, 1993.) This latter set of experiments suggests that aside from being concerned with their own monetary payoffs, subjects appear to be concerned also with the monetary payoffs of others. Preferences having this property are commonly referred to as *interdependent*.

For any specific experimental environment, it is usually possible to find plausible specifications of preference interdependence that fit the data. For instance, altruistic preferences can explain contributions in public goods environments, and an envious concern for relative payoffs is consistent with data from bargaining games (Andreoni and Miller, 1999, Bolton, 1991). The challenge facing those who attempt to provide a parsimonious alternative to the hypothesis of material self-interest is that a single specification should simultaneously explain a wide variety of experimental results. A number of recent attempts to meet this challenge have been made (Rabin, 1993, Fehr and Schmidt, 1997, Bolton and Ockenfels, 1997, Levine, 1998, Falk and Fischbacher, 1998, and Dufwenberg and Kirchsteiger, 1998.) As a result, there are now a variety of competing specifications of preference interdependence, each of which is consistent with results from several experiments. What remains to be determined, however, is whether any of these specifications can be provided with a convincing evolutionary rationale. This raises the question of how particular forms of preference interdependence may have emerged and persisted in human societies.

In this paper, we provide an evolutionary account of the emergence and stability of *reciprocal preferences* of the kind that Levine (1998) has used to confront the experimental data. Individuals endowed with such preferences are concerned not only with their own material payoffs but also with the material payoffs of others. This concern may be altruistic or spiteful and is represented by (positive or negative) weights placed on the payoffs of others. These weights themselves vary systematically with the degree of altruism or spite that others are perceived to possess, so that the well-being of a fellow altruist is given greater weight by an altruist than is the well-being of a selfish or spiteful individual. Such preferences have two quite distinct evolutionary advantages. First, groups consisting largely or entirely of individuals endowed with such preferences behave in a manner similar to groups of altruists, which makes them more efficient than groups of self-regarding materialists in many strategic environments. This tends to strengthen their prospects for survival when there is some competition among groups. Second, unlike pure altruists, individuals who are endowed with reciprocal preferences can survive and spread *within* groups consisting largely of self-regarding materialists. This occurs because the predominance of materialists in the group can induce reciprocators to act as if they had spiteful preferences, and the possession of spiteful preferences is known to yield a strategic advantage over materialists in a variety of environments (Koçkesen *et al.* 1997, 1998). One such class of environments consists of aggregative games, which possess the property that an individual's material payoff depends only on her own action and an aggregate of the actions of others. Although such a payoff structure has usually been associated with strategic market games (Dubey *et al.*, 1980, Corchón, 1996), it also includes, for instance, common pool resource extraction and public goods games. Such environments have been important in human interaction from the earliest times and remain economically important to this day. The focus of this paper is accordingly on aggregative games.

The flexibility in behavior that reciprocal preferences provide makes them effective under both individual selection (when they are outnumbered) and under group selection (when they are prevalent). By the same token, such preferences may suffer from two evolutionary disadvantages. In groups consisting largely of reciprocators, materialists can thrive when they are few in number, provided that the possibility of sanctioning specific individuals is precluded. This occurs because reciprocators continue to act altruistically, unwilling to reduce the material well-being of their fellow reciprocators in order to sanction the few materialists in their midst. Similarly, the spiteful behavior of reciprocators in groups consisting predominantly of materialists lowers average group fitness, so their advantage under individual selection may be reduced when some measure of group selection is also at work.

The net effect of all these considerations is difficult to gauge without a formal model of the strategic environment and the selection process. Such a model is provided below, and has the following features. Interaction is strategic and occurs in finite groups. Preferences may be heterogenous within a group, with some individuals pursing their material self-interest, while others have reciprocal preferences. Individuals behave rationally given their preferences and are assumed to take actions consistent with an equilibrium of the game. Individuals with different preferences will typically take different equilibrium actions and receive different payoffs, and it is this payoff differential which drives the evolutionary dynamics. Under individual selection, we show that a group consisting exclusively of materialists will not generally be stable in the presence of reciprocal preferences. We next allow for random (nonassortative) matching: at the end of each period of interaction, all individuals are randomly matched with others in the global population to form new groups. Again, it is shown that a population of materialists will not generally be stable in the presence of reciprocators, although the class of preferences that can invade a materialist population is restricted relative to the case of individual selection. Finally, we examine the efficiency effects of changes in group composition and show that reciprocal preferences are efficiency-reducing when they are rare, and efficiency-enhancing when they are widespread. This suggests that such preferences can thrive under assortative matching and group selection. In comparison with simpler specifications of preference interdependence (such as pure altruism or envy), therefore, the survival of reciprocal preferences is less sensitive to details of the evolutionary selection process.

Our approach to the evolution of preferences differs from earlier work in one or more of the following four respects. First, the manner in which preference interdependence is conceived is context-free, and can therefore be applied to arbitrary strategic environments. This may be contrasted with work in which the formalization of preference interdependence is tailored to specific environments, such as bargaining games or team production (Güth and Yaari, 1992, Bowles and Gintis, 1998). It may also be contrasted with the recent work of Ely and Yilankaya (1997) and Dekel *et al.* (1998) in which the class of preferences is composed of orderings over action profiles for given games rather than orderings over material payoff profiles which could apply to any game. Second, we examine the evolution of preferences in the class of aggregative games rather than in the context of a particular economic model. Our results therefore apply to all models belonging to this class, of which there are several. In these two respects, the methodology adopted here follows Bester and Güth (1998), who examine conditions for the survival of altruistic preferences under pairwise random matching, and Koçkesen *et al.* (1997, 1998), who focus on the survival of envious preferences under individual selection. Third, we consider the evolutionary prospects of a class of preferences that have already been reasonably successful in explaining data from a variety of experimental settings (Levine, 1998). And fourth, we consider the evolution of such preferences in a each of the most commonly studied evolutionary environments: individual selection, random matching, assortative interaction and group selection.[1]

[1] Several papers from a variety of disciplines have addressed the question of the evolution of reciprocity as an aspect of *behavior* rather than an attribute of preferences. The seminal contributions of Trivers (1971) and Axelrod and Hamilton (1981) demonstrated that direct reciprocity (interpreted as the repayment of an altruistic act) can emerge when interactions are repeated. However, such behavior is known to be consistent

2 Preference Interdependence

Let $\Gamma \equiv \{X_i, \pi_i\}_{i \in I}$ be an n–person normal form game where $I = \{1, ..., n\}$ is the set of players, X_i denotes the action set of player i and $\pi_i : \times_j X_j \rightarrow \mathbf{R}$, $i \in I$, the *material* payoff functions. In the context of experimental games, material payoffs correspond to cash payments. More generally, material payoffs may be interpreted to be any magnitude, such as income, wealth, or fitness, for which interpersonal comparisons are possible. If preferences are *independent* (an individual's ranking of payoff profiles depends only on their own material payoff), then the game Γ provides a complete description of the strategic interaction in which the players are engaged. If, on the other hand, preferences are *interdependent*, then the utility $u_i : \times_j X_j \rightarrow \mathbf{R}$ of player i will depend on the entire distribution of material payoffs resulting from any given action profile $x \in \times_j X_j$. We may write

$$u_i(x) = F_i(\pi_1(x), ..., \pi_n(x)).$$

If F_i is strictly increasing in π_j for all $j \neq i$, individual i is an altruist; if it is strictly decreasing then i has envious or spiteful preferences. Altruistic preferences have been argued to underlie behavior in public goods and dictator game experiments, while envious preferences have been advanced to explain data from bargaining experiments. The problem with pure altruism or envy, however, is that while each is consistent with data from some environments, both are flatly contradicted by others. More complex forms of preference interdependence are therefore required if data from a variety of experiments is to be simultaneously confronted.

Several recent papers have attempted to meet this challenge. These papers fall into three broad categories. Fehr and Schmidt (1997) and Bolton and Ockenfels (1997) provide specifications of preference interdependence that are *object-oriented,* in that individuals are assumed to care only about the distribution of material payoffs and not about the intentions or preferences of those with whom they interact. Although they differ with respect to a number of details, both papers require that individuals experience some disutility from being at either extreme of the payoff distribution. These papers are able to explain much more of the data than can simpler specifications of preference interdependence, but cannot account

with materialist preferences since the prospect of future gain or the credible threat of future punishment can induce sufficiently patient selfish individuals to make material sacrifices in repeated games (see, for instance, Fudenberg and Maskin, 1986). More recently, Boyd and Richerson (1989) and Nowak and Sigmund (1998) have explored conditions for the evolution of *indirect reciprocity*, interpreted as a willingness to sanction antisocial behavior even by those who were not victims of that behavior. As in the earlier literature on direct reciprocity, behavior in this case is not preference based, and it is unclear whether or not the practice of indirect reciprocity is also consistent with materialist preferences.

for the fact that at least in some environments, subjects consistently choose very different payoff distributions depending on whether they were generated by a random device or by the intentional actions of other players (see Blount, 1995, for one such set of experiments, and Falk and Fischbacher, 1998, for additional references.)

A second group of papers adopts the approach of *psychological games* in which player utilities depend not just on action profiles but also on their initial beliefs (Rabin, 1993, Dufwenberg and Kirchsteiger, 1998, Falk and Fischbacher, 1998).[2] In equilibrium, all beliefs (including higher-order beliefs) are correct, and individuals take optimal actions conditional on these beliefs and the actions of others. Different beliefs (corresponding to different equilibria) imply possibly different utility profiles at any given action profile. This endogeneity of utility profiles represents a considerable departure from standard game theoretic methodology. Papers using the apparatus of psychological games to explain data from experiments are based on the hypothesis that beliefs about the kindness or unkindness of opponent strategies will give rise to the desire to reciprocate, where the kindness or unkindness of an individual's strategy is assessed in terms of the (material) payoff implications of other strategies available to them. These papers are effective in accounting for the role of intentionality in experimental results. As presently formulated, however, they cannot account for the fact that extraneous information that directly pertains to an opponent's preferences (such as the opponent's behavior in some prior experiment with other subjects) influences the manner in which they are treated, for instance, in dictator games.[3]

A third approach, which applies the standard game theoretic methodology, is based on the hypothesis of *reciprocal preferences*. Here an individual's utility is directly influenced by parameters that enter the utility functions of others. This approach, due to Levine (1998), can account for both intentionality and the importance of extraneous information pertaining to the preferences of others. Intentionality is important because intentional actions can reveal information about underlying preferences, but any source of such information, such as past behavior in other experiments, will influence the manner in which a person is treated.

Levine suggests the following specification of reciprocal preferences, which allow for both

[2] See Geanokopolous *et al.* (1989) for the methodology and fundamental properties of psychological games.

[3] As an example, consider the results of Kahneman *et al.* (1986), who asked subjects whether they wished to share $12 equally with an opponent who had made an unequal proposal in a prior ultimatum bargaining experiment or $10 equally with one who had made an equal proposal. In either case, the opponent *not* selected to receive a share would receive nothing. Almost three-quarters of subjects chose the latter option, indicating a willingness to reward kindness (and punish unkindness) even when it had been directed at others.

altruism and spite:

$$u_i(x) = \pi_i(x) + \sum_{j \neq i} \beta_{ij} \pi_j(x), \tag{1}$$

where

$$\beta_{ij} = \frac{\alpha_i + \lambda_i \alpha_j}{1 + \lambda_i}$$

and $-1 < \alpha_i < 1$ and $0 \leq \lambda_i$. Here α_i may be interpreted as a measure of an individual's pure altruism, and λ_i a measure of the degree to which the weight β_{ij} placed by individual i on the material payoffs of individual j is sensitive to the altruism of the latter. Levine argues that a suitably chosen, stable distribution of preferences belonging to this class can simultaneously account for results from ultimatum bargaining, competitive auction, centipede, and public goods games. Note, however, that in Levine's specification an individual i with $\alpha_i > 0$ can never place a negative weight on the payoffs of an individual j who is purely self-interested ($\alpha_i > 0$ and $\alpha_j = 0$ implies $\beta_{ij} > 0$). Such "flexible altruists" would be driven to extinction under individual selection in the class of strategic environments considered here. The following slight variant of Levine's specification, however, has considerably greater prospects for survival under evolutionary pressure:

$$\beta_{ij} = \frac{\alpha_i + \lambda_i (\alpha_j - \alpha_i)}{1 + \lambda_i}. \tag{2}$$

In this case, it is not the extent of the other party's altruism that counts, but rather the deviation of their altruism from one's own. A purely self-interested individual i, whose only concern is with her own material payoffs corresponds to the case $\alpha_i = \lambda_i = 0$. We shall refer to such individuals as materialists. A pure altruist who puts the same positive weight on the payoffs of all others is represented by $\alpha_i > 0 = \lambda_i$. A player with $\alpha_i = 0 < \lambda_i$ places no weight on the payoffs of a self-interested person but places positive weight on the payoffs of pure altruists. More generally, if $\alpha > 0$ and $\lambda > 0$, an individual is altruistic towards those who are similarly inclined but is also capable of being spiteful toward materialists. We shall refer to those with this preference simply as reciprocators. It is assumed that $0 \leq \alpha_i < 1$ and $\lambda_i \geq 0$. These two conditions ensure that $-1 < \beta_{ij} < 1$, so that each person places more weight on their own material payoff than on that of another. It is argued below that preferences belonging to this class can invade a population of purely self-interested individuals under a variety of evolutionary selection mechanisms.[4]

[4] Since the specification of reciprocal preferences used here is very similar in structure to one used by Levine, and has the same number of free parameters, one might reasonably expect that there will be negligible difference in the degree to which the two specifications can account for experimental regularities. From an evolutionary perspective however, the two specifications differ sharply in their prospects for survival.

3 Strategic Advantage

Central to the following analysis of preference evolution within a group is the idea of *strategic advantage*. Informally stated, a given preference yields a strategic advantage over another if a player having the former preference obtains a greater payoff than an otherwise identical player with the latter preference at any equilibrium. Although the concept of strategic advantage may be applied to arbitrary games in which two or more players are symmetrically placed with respect to each other, it is particularly useful in the context of symmetric games (since all players are "otherwise identical" in this case.) In this paper attention is confined to symmetric games.

Let $\Gamma \equiv \{X, \pi_i\}_{i \in I}$ be a symmetric n–person normal form game, where X denotes the (common) action space and the material payoffs π_i are symmetric ($\pi_i = f(x_i, x_{-i})$ for some function f which is common to all players.) Although each of the players has the same action space and the same material payoff function, players may differ with respect to their preferences. Suppose that some subset M of the players have materialist preferences, so that $u_i(x) = \pi_i$ for all $i \in M$. The set of remaining players R have reciprocal preferences, so that $u_i(x)$ is given by (1–2) for all $i \in R$. The resulting strategic interaction is then described by an (asymmetric) n–person normal from game in which each player's action space is X and the objective functions are as given above. Let $\Gamma(k)$ denote this game, where $k \in \{0, ..., n\}$ is the number of players with materialist preferences. The idea of strategic advantage can then be expressed as follows. We say that reciprocal preferences yield a *strategic advantage* over materialists preferences at the preference distribution k if, at each Nash equilibrium x of $\Gamma(k)$,

$$\pi_i(x) \geq \pi_j(x) \qquad \text{for all } (i, j) \in R \times M$$

with strict inequality holding for some (i, j). If the above inequality is reversed (and holds strictly for some $(i, j) \in R \times M$) then materialist preferences yield a strategic advantage over reciprocal preferences. In any group in which some preference yields a strategic advantage over another, the population share of the former will tend to increase relative to that of the latter, assuming that the dynamics of the population composition are payoff monotonic with respect to material payoffs. The requirements for strategic advantage are stringent, however, since the advantage must exist at *each* Nash equilibrium of $\Gamma(k)$. Hence it may commonly be the case that neither preference yields a strategic advantage over the other. One way to make the concept of strategic advantage operational, therefore, is to further restrict the class of games considered.

We restrict attention to games in which the action space $X = [a, b] \subset \mathbf{R}$ is a closed interval, and in which the material payoff functions are of the form

$$\pi_i(x) = H(x_i, n\bar{x}), \tag{3}$$

where $n\bar{x} = \sum_{j=1}^{n} x_j$ is the aggregate action in the group, and H is assumed to be twice differentiable. This is the class of symmetric *aggregative games* (Dubey *et al.*, 1980) which we denote by $\mathcal{A}$. Let $T(x_i, n\bar{x})$ denote the marginal payoff of player i:

$$\frac{\partial \pi_i}{\partial x_i} = H_1(x_i, n\bar{x}) + H_2(x_i, n\bar{x}) \equiv T(x_i, n\bar{x}).$$

Most of the results below are based on one or more of the following additional restrictions on the payoff functions:

$$H_1 > 0 \tag{AM}$$

$$H_2 < 0 \tag{NS}$$

$$H_{11} + H_{21} = T_1 < 0 \tag{A1}$$

$$H_{21} + H_{22} = T_2 < 0 \tag{SS}$$

The first of these is the assumption of (positive) *action monotonicity*: at any given action profile, a player with a higher action obtains a higher payoff. The second is the assumption of negative spillovers, and implies that an increase in the action of one player lowers the payoffs of all others.[5] Assumptions (A1) and (SS) state that the marginal payoff function $T(x_i, n\bar{x})$ is strictly decreasing in both components (the latter corresponds to the assumption of strategic substitutability.) Note that (A1) and (SS) together imply strict concavity of payoffs in own actions. Both these assumptions are common in analyses of aggregative games (see, for instance, Corchón 1996), together with the following conditions, which are made to exclude boundary equilibria of limited interest in the present context.

$$T(a, na) > 0 > T(b, nb). \tag{BC}$$

The following examples illustrate that the above assumptions can indeed be satisfied by games which have economically meaningful interpretations and which are relevant environments in which the question of preference evolution may be examined.

[5] Note that any symmetric game satisfying negative action monotonicity and positive spillovers, by a suitable relabeling of actions, can be transformed into one which satisfies positive action monotonicity and negative spillovers. Hence all results which use (AM–NS) continue to hold if the signs of *both* inequalities are reversed. See Koçkesen *et al.* (1997, 1998) for further discussion and applications of these conditions to the analysis of strategic advantage in the case of envious or spiteful preferences.

Example 1. (Private Provision of Public Goods). Suppose each of n individuals has an endowment b of a private good, part or all of which can be contributed towards the provision of a public good. Individual i's action x_i is the amount of the good retained for private use. The aggregate contribution to the public good is then $nb - n\bar{x}$. The action space of each player is $[0, b]$ and the material payoff functions are $\pi_i = H(x_i, n\bar{x}) = f(x_i) + g(nb - n\bar{x})$ where $f', g' > 0$ and $f'', g'' < 0$. This game is aggregative and satisfies (AM), (NS), (A1) and (SS). If, in addition, $f'(0) > g'(nb)$ and $f'(nb) < g'(0)$, it satisfies (BC).

Example 2. (Common Pool Resource Extraction). Suppose each of n individuals has access to a common pool resource. Let $x_i \geq 0$ denote the extraction effort of individual i, and $n\bar{x}$ the aggregate extraction effort. Total output of the resource is given by the production function $f(n\bar{x})$, assumed to satisfy $f(0) = 0$, $f'(0) > w$, and $f'' < 0$, where w is a constant average cost of extraction effort. Let $A(n\bar{x}) = f(n\bar{x})/(n\bar{x})$ denote average extraction per unit of effort and set $A(0) = \lim_{\bar{x}\to 0} f(n\bar{x})/n\bar{x} = f'(0)$. Concavity of f implies that $A' < 0$. The material payoff obtained by each extractor is proportional to her extraction effort and is given by $\pi_i(x) = H(x_i, n\bar{x}) = x_i(A(n\bar{x}) - w)$. If $nA' + n\bar{x}A'' < 0$, it can be shown that there exists a set $[a, b]^n$ in which all equilibrium action profiles must lie and which has the following property: restricting the action set of this game to $[a, b]$ yields an aggregative game which satisfies (AM), (NS), (A1), (SS), and (BC).[6]

As a special case to be considered below, we say that a game $\Gamma \in \mathcal{A}$ is *separable* if $H_{12} = 0$. In this case material payoffs may be expressed as the sum of two separate functions of x_i and $n\bar{x}$ respectively. Note that the game described in Example 1 is separable, but that in Example 2 is not.

4 Individual Selection

Consider any game $\Gamma \in \mathcal{A}$ and suppose that $k \leq n$ players pursue the maximization of their material payoffs, while the remainder have preferences described by (1-2) for some value of $\alpha \in (0, 1)$ and $\lambda > 0$. We shall refer to individuals of the former type as materialists, and the latter as reciprocators. The objective functions are then given by

$$u_i(x) = \begin{cases} \pi_i(x) & \text{for all } i \in M, \\ \pi_i(x) + \beta_r \sum_{j \in R\setminus\{i\}} \pi_j(x) + \beta_m \sum_{j\in M} \pi_j(x) & \text{for all } i \in R, \end{cases} \tag{4}$$

[6]Since $A' < 0$, the assumption $nA' + n\bar{x}A'' < 0$ is trivially satisfied if $A'' \leq 0$. It is also satisfied if the production function is of the form $f(n\bar{x}) = (n\bar{x})^{\theta}$ where $\theta \in (0, 1)$, as is commonly assumed in this context.

where R is the set of reciprocators, M the set of materialists, and β_r and β_m satisfy

$$\beta_r = \frac{\alpha}{1+\lambda}, \qquad \beta_m = \frac{\alpha(1-\lambda)}{1+\lambda}.$$

Since $\lambda > 0$, $\beta_r \in (0, \alpha)$ and $\beta_m \in (-\alpha, \alpha)$. Reciprocators are spiteful towards materialists if $\lambda > 1$, and altruistic towards all players if $\lambda < 1$. Assume for the moment that the distribution of preferences is common knowledge though the particular assignment of preferences to individuals need not be known. The set of objective functions (4), together with the population composition k, the action space $[a, b]$ available to each player and the material payoff functions (3) define the n–person game $\Gamma(k)$. Although the material payoff functions are symmetric, equilibria of $\Gamma(k)$ will generally be asymmetric whenever there is heterogeneity with respect to player objective functions.

Under individual selection, assuming that the preference distribution evolves according to dynamics that are monotonic in material payoffs, k will increase or decrease depending on which of the two preferences yields greater equilibrium payoffs in $\Gamma(k)$. In accordance with this hypothesis, we say that the state $k = n$ is *stable* if materialist preferences yield a strategic advantage over reciprocal preferences in $\Gamma(n-1)$, and *unstable* if the reverse is true. Analogously, the state $k = 0$ is stable if reciprocal preferences yield a strategic advantage over materialist preferences in $\Gamma(1)$, and unstable if the reverse is true. In other words, a monomorphic state is stable if the appearance of a single mutant alters the equilibrium set in such a way as to yield the mutant a lower payoff than each player with the incumbent preference, and a strictly lower payoff than some such player. The following result establishes conditions under which a group of materialists is unstable under individual selection in the presence of reciprocators.

Proposition 1. *Consider any* $\Gamma \in \mathcal{A}$ *which satisfies* (AM), (NS), (A1), *and* (BC). *The state* $k = n$ *is unstable if* $\lambda > 1$ *and stable if* $\lambda < 1$.

Proof. Let $u_i = \pi_i$ for all $i \neq n$ and $u_n = \pi_n + \beta_m \sum_{j \neq n} \pi_j$. If $\lambda > 1$, then $\beta_m < 0$. Let x be any Nash equilibrium of $\Gamma(n-1)$. We claim that $x = (y, ..., y, z)$ for some $y, z \in [a, b]$. To see this, suppose there exist $i, j \in M$ such that $x_i > x_j$. Then a necessary condition for equilibrium is $T(x_i, n\bar{x}) \geq T(x_j, n\bar{x})$. This implies $x_i \leq x_j$ from (A1), a contradiction. Hence $x = (y, ..., y, z)$ for some $y, z \in [a, b]$. If $z = b$ or $y = a$, then $z > y$ from (BC), and so $\pi_n > \pi_i$ for all $i \in M$ from (AM). The state $k = n$ is therefore unstable in this case. Now suppose $z < b$ and $y > a$. In this case $T(y, n\bar{x}) \geq 0$ and

$$\frac{\partial u_n}{\partial x_n} = T(z, n\bar{x}) + \beta_m (n-1) H_2(y, n\bar{x}) \leq 0.$$

Since $H_2 < 0$ from (NS) and $\beta_m < 0$, this implies $T(z, n\bar{x}) < 0 \leq T(y, n\bar{x})$. Hence $z > y$ from (A1), and so $\pi_n > \pi_i$ for all $i \in M$ from (AM). The state $k = n$ is therefore unstable in this case too. A similar argument can be used to show that if $\lambda < 1$, then $\pi_n < \pi_i$ for all $i \in M$, so that $k = n$ is stable. ■

Proposition 1 implies that a population of materialists is vulnerable to invasion by reciprocators, provided that $\lambda > 1$ and the selection dynamics are monotonic in material payoffs. The fact that $\lambda > 1$ is necessary in order for a preference distribution in which all players are materialists to be unstable is intuitive, since reciprocators are altruistic even towards materialists when $\lambda < 1$. It is the potentially spiteful behavior of reciprocators which gives them the advantage over materialists.

This result does not depend on the complete information about the distribution of preferences that has been assumed. We may assume instead that all players assign sufficiently high probability to materialists being materialists. For expositional simplicity, consider the case in which the group consists of only two players. Suppose that players share a common prior that assigns probability p to any given player being a materialist. They then become perfectly informed of their own preferences and receive no information concerning the preferences of their opponent. This defines a Bayesian game in which there are two types of each player (a materialist type and a reciprocator type.) Let x_{mi} and x_{ri} respectively be the actions taken by a materialist player i and a reciprocator player i in this game. The expected payoffs of the two types of each player are given by the following, where $i, j \in \{1, 2\}$ and $i \neq j$.

$$v_{mi}(x_{mi}, x_{mj}, x_{ri}, x_{rj}) = \rho H\left(x_{mi}, x_{mi} + x_{mj}\right) + (1 - \rho) H\left(x_{mi}, x_{mi} + x_{rj}\right)$$

and

$$\begin{aligned} v_{ri}(x_{mi}, x_{mj}, x_{ri}, x_{rj}) = \rho &\left(H\left(x_{ri}, x_{mj} + x_{ri}\right) + \beta_m H\left(x_{mj}, x_{mj} + x_{ri}\right)\right) \\ &+ (1 - \rho)\left(H\left(x_{ri}, x_{ri} + x_{rj}\right) + \beta_r H\left(x_{rj}, x_{ri} + x_{rj}\right)\right) \end{aligned}$$

By continuity of objective functions in probabilities and actions, the correspondence from probabilities to Nash equilibrium actions is upper hemi-continuous, so that for p sufficiently close to 1, any equilibrium of this game must lie close to an equilibrium of the game in which $p = 1$ (that is, each player believes with probability 1 that all other players are materialists.) If it can be shown that when $p = 1$, a reciprocator obtains a higher (lower) payoff than a materialist if $\lambda >$ ($<$) 1, then this would be true also when the common prior probability that each player is materialist is sufficiently close to (but less than) 1. Accordingly, suppose that each player believes with probability 1 that all other players are materialists. Since $\Gamma(n)$ has a unique equilibrium under (A1) and (SS) (see Corchón, 1996, Proposition 1.3), and we are considering the case $n = 2$, each materialist will play c, where (c, c) is the unique equilibrium of $\Gamma(2)$. From (BC), $c \in (a, b)$ so $T(c, 2c) = 0$. Now suppose that reciprocators choose the action d in equilibrium. Note that

$$\frac{\partial v_{ri}}{\partial x_{ri}}(1, c, d) = T(d, c + d) + \beta_m(n - 1)H_2(c, c + d).$$

If $d \leq c$ then from (A1), $T(d, c+d) \geq 0$. By (NS), $H_2(c, c+d) < 0$. So if $\beta_m < 0$, which is the case when $\lambda > 1$, then $\partial v_{ri}/\partial x_{ri} > 0$ in equilibrium, implying $d = b > c$, a contradiction. So when $\lambda > 1$, the optimal action for a reciprocator must be greater than c, the optimal action for a materialist. By analogous reasoning, when $\lambda < 1$, the optimal action for a reciprocator must be less than the optimal action for a materialist. Hence the conclusion of Proposition 1 holds even when the reciprocator is not only *not* recognized, but is believed, with sufficiently high probability, to be a materialist.

Returning to the complete information model, notice that as the population share of reciprocators rises, their actions become increasingly altruistic since there are fewer materialists in their presence. Hence, for a range of parameter values, materialists will not be eliminated entirely. In order to state this formally, the following definition is helpful. An equilibrium x of $\Gamma(k)$ is said to be *intragroup symmetric* if players with the same preference take the same action. Formally, x is intragroup symmetric if $x_i = x_j$ if either $i, j \in M$ or $i, j \in R$. The following lemma identifies conditions under which equilibria of $\Gamma(k)$ are intragroup symmetric for all k.

Lemma 1. *Suppose* $\Gamma \in \mathcal{A}$ *satisfies* (A1) *and* $H_{12} \geq 0$. *Then, for any* $k \in \{0, ..., n\}$, *and any equilibrium* x *of* $\Gamma(k)$ *is intragroup symmetric.*

Proof. Suppose there exist $i, j \in R$ such that $x_i > x_j$ at some equilibrium x of $\Gamma(k)$. Then we must have $\partial u_i/\partial x_i \geq 0 \geq \partial u_j/\partial x_j$, or

$$T(x_i, n\bar{x}) + \beta_r H_2(x_j, n\bar{x}) + \beta_r \sum_{j \in R \setminus \{i,j\}} H_2(x_j, n\bar{x}) + \beta_m \sum_{j \in M} H_2(x_j, n\bar{x}) \geq$$
$$T(x_j, n\bar{x}) + \beta_r H_2(x_i, n\bar{x}) + \beta_r \sum_{j \in R \setminus \{i,j\}} H_2(x_j, n\bar{x}) + \beta_m \sum_{j \in M} H_2(x_j, n\bar{x}).$$

where $\beta_r > 0$. This implies $T(x_i, n\bar{x}) + \beta_r H_2(x_j, n\bar{x}) \geq T(x_j, n\bar{x}) + \beta_r H_2(x_i, n\bar{x})$. Since $H_{12} \geq 0$, $H_2(x_i, n\bar{x}) \geq H_2(x_j, n\bar{x})$. Hence $T(x_i, n\bar{x}) \geq T(x_j, n\bar{x})$. This implies $x_i \leq x_j$ from

(A1) , a contradiction. The proof that $x_i = x_j$ for all $i, j \in M$ follows by setting $\beta_r = \beta_m = 0$ and applying the above reasoning. (Note that $H_{12} \geq 0$ is not required in this case.) ■

The following result identifies the relevant parameter range for which a population of reciprocators will be vulnerable to invasion by materialists in the special case of separable payoff functions.

Proposition 2. *Consider any separable* $\Gamma \in \mathcal{A}$ *which satisfies* (AM), (NS), (A1), *and* (BC). *The state* $k = 0$ *is unstable if* $\lambda < n - 1$ *and stable if* $\lambda > n - 1$.

Proof. Let $u_1 = \pi_1$ and $u_i = \pi_i + \beta_r \sum_{j \in R \setminus \{i\}} \pi_j + \beta_m \pi_1$ for all $i \neq 1$. Let x be any Nash equilibrium of $\Gamma(1)$. From Lemma 1, $x = (y, z, ..., z)$ for some $y, z \in [a, b]$. Suppose first that $\lambda < n-1$. If $z = a$ or $y = b$, then $y > z$ from (BC) , and so $\pi_1 > \pi_i$ for all $i \in R$ from (AM) . Hence $k = 0$ is unstable in this case. Now suppose $y < b$ and $z > a$. Then $T(y, n\bar{x}) \leq 0$ and

$$\frac{\partial u_n}{\partial x_n} = T(z, n\bar{x}) + \beta_r (n-2) H_2(z, n\bar{x}) + \beta_m H_2(y, n\bar{x}) \geq 0.$$

Since $H_{12} = 0$, $H_2(y, n\bar{x}) = H_2(z, n\bar{x})$ so we have

$$T(z, n\bar{x}) + (\beta_r (n-2) + \beta_m) H_2(z, n\bar{x}) \geq 0.$$

Since $\lambda < n - 1$, $\beta_r (n-2) + \beta_m = (n - 1 - \lambda) \alpha / (1 + \lambda) > 0$. This, together with (NS) and the above relation yields $T(z, n\bar{x}) > 0 \geq T(y, n\bar{x})$. Hence $z < y$ from (A1), and $\pi_1 > \pi_i$ for all $i \in R$ from (AM) , so $k = 0$ is unstable. A similar argument can be used to show that if $\lambda > n - 1$, then $\pi_1 < \pi_i$ for all $i \in R$, so $k = 0$ is stable. ■

For the class of games to which the above result applies, a single materialist in a group of reciprocators will outperform all others at the equilibrium action profile, as will a single reciprocator in a group of materialists, provided that $1 < \lambda < n - 1$. If the population composition evolves under pressure of differential material payoffs, neither monomorphic state will be stable, and the population will be polymorphic in the long run. The intuition underlying this is the following. In a population of materialists, a single reciprocator places negative weight on the payoffs of all others. Relative to an equilibrium in which all players are materialists, the reciprocator is tempted to increase her action despite the fact that this increase reduces her material payoff, since it reduces the payoffs of materialists. This increase lowers the marginal returns to an increase in action for all players, and induces the materialists to respond by reducing their action. Although the overall effect may be to reduce the average material payoff in the group as a whole, the reciprocator outperforms the materialists since his equilibrium action is higher. On the other hand, a materialist can thrive in a population of reciprocators, provided that their altruism towards each other prevents them from raising their actions for punitive purposes when a single materialist is in their midst. The necessary condition for this to occur is that $\lambda < n-1$. If this inequality is reversed, the negative weight placed by reciprocators on the payoffs of materialists is so great that it outweighs the effects of their mutual altruism, and a monomorphic group of reciprocators is stable in this case.

5 Random Matching

The previous section considered the dynamics of the preference distribution within a single group. We now turn to the question of the long-run preference distribution in a large population, the members of which are matched randomly with each other in groups. For convenience, we assume that the population is infinite, though our results continue to hold for populations that are sufficiently large.

Let p denote the share of materialists in the global population. The probability $\gamma_k(p)$ that a randomly selected group will contain k materialists is then

$$\gamma_k(p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

As before, assume that within a group the distribution of preferences is common knowledge and that the members of the group are able to locate an equilibrium of the game (the case of incomplete information is discussed below.) Let $\mu_m(k)$ denote the expected payoff to materialists in groups with population composition k, and let $\mu_r(k)$ be the expected payoff to reciprocators. In the presence of multiple equilibria, these payoffs will depend on the probabilities with which the various equilibria are realized. For the results to follow, it is irrelevant which equilibria are realized and in what proportions. We therefore assume that for any given population composition k,there is some exogenously given probability that any particular equilibrium will be realized, so that $\mu_m(k)$ and $\mu_r(k)$ are well defined. Let $\bar{\mu}_m(p) = \sum_{k=1}^{n} \gamma_k(p)\mu_m(k)$ be the expected payoff to materialists in the population as a whole, with $\bar{\mu}_r(p) = \sum_{k=0}^{n-1} \gamma_k(p)\mu_r(k)$ being the corresponding expected payoff to reciprocators. We are interested in the stability of the states $p = 0$ and $p = 1$ under payoff monotonic selection dynamics. Payoff monotonicity here corresponds to the assumption that for all $p \in (0,1)$,the following holds:

$$\bar{\mu}_m(p) > (<)\ \bar{\mu}_r(p) \Leftrightarrow \dot{p} > (<)\ 0,$$

with $\dot{p} = 0$ for $p \in \{0, 1\}$. With an infinite global population, a sufficient condition for the instability of the state $p = 1$ is that $\mu_m(n) < \mu_r(n-1)$. This follows from the fact that $\lim_{p \to 1} \gamma_n(p) = 1$ (so that almost all materialists will be in monomorphic groups when p is close to 1) and $\lim_{p \to 1} \gamma_{n-1}(p) / \sum_{k=0}^{n-1} \gamma_k(p) = 1$ (so that almost all reciprocators will be in groups in which all other players are materialists when p is close to 1.) Similarly, a sufficient condition for the stability of the state $p = 0$ is that $\mu_m(1) < \mu_r(0)$. The following result identifies conditions under which reciprocators can invade a population of materialists under random matching.

Proposition 3. *Consider any* $\Gamma \in \mathcal{A}$ *which satisfies* (AM), (NS), (A1), (SS), *and* (BC). *There exists* $\bar{\lambda} > 1$ *such that if* $1 < \lambda < \bar{\lambda}$, *the state* $p = 1$ *is unstable.*

Proof. Define $G(y, z) \equiv T(y, (n-1)y + z)$ and note that from (A1) and (SS), $G_1 = T_1 + (n-1)T_2 < 0$ and $G_2 = T_2 < 0$. By the implicit function theorem, $G(y, z) = 0$ defines a differentiable function $y = b(z)$ such that $G(b(z), z) = 0$ and $b'(z) = -G_2/G_1 < 0$. Recall (from the proof of Proposition 1 above) that all equilibria of $\Gamma(n-1)$ are of the form $(y, ..., y, z)$ Hence $G(y, z) = 0$ and $y = b(z)$ must hold at any equilibrium $(y, ..., y, z)$ of $\Gamma(n-1)$ at which $y \in (a, b)$.

Under (A1) and (SS), $\Gamma(n)$ has a unique equilibrium (Corchón, 1996, Proposition 1.3), which must therefore be symmetric. Let $(c, ..., c)$ denote this equilibrium. From (BC), $c \in (a, b)$. Hence $T(c, nc) = G(c, c) = 0$ and $b(c) = c$. Define $\varphi(z) \equiv H(z, (n-1)b(z) + z)$ and note that $\varphi'(c) = H_1(c, nc) + ((n-1)b'(c) + 1)\, H_2(c, nc) = T(c, nc) + (n-1)b'(c)H_2(c, nc) = (n-1)b'(c)H_2(c, nc) > 0$ since $T(c, nc) = 0$, $H_2 < 0$ from (NS) and $b' < 0$. This implies that there exists $\bar{\varepsilon} > 0$ such that $\varphi(c + \varepsilon) > \varphi(c)$ for all $\varepsilon \in (0, \bar{\varepsilon}]$. Note that $\pi_n(y, ..., y, z) = \varphi(z)$ at any equilibrium $(y, ..., y, z)$ of $\Gamma(n-1)$ at which $y \in (a, b)$.

With $\alpha > 0$ given, let $E(\lambda)$ denote the set of equilibria of $\Gamma(n-1)$ when the reciprocator has preference parameter $\lambda \geq 1$. Consider a sequence $\left(\lambda^t\right)_{t=0}^{\infty}$ where $1 < \lambda^t < \lambda^{t-1}$ for all $t \geq 1$ and $\lim_{t \to \infty} \lambda^t = 1$. Since the Nash equilibrium correspondence has a closed graph, any sequence $(y^t, ..., y^t, z^t)_{t=0}^{\infty}$ with $(y^t, ..., y^t, z^t) \in E(\lambda^t)$ has a limit point in $E(1)$. Note that when $\lambda = 1$, $\Gamma(n)$ and $\Gamma(n-1)$ are identical games so $\Gamma(n-1)$ also has a unique interior equilibrium at $(c, ..., c)$. Hence $E(1)$ consists of the single element $(c, ..., c)$ and any sequence $(y^t, ..., y^t, z^t)_{t=0}^{\infty}$ with $(y^t, ..., y^t, z^t) \in E(\lambda^t)$ converges to $(c, ..., c)$.

We claim that for any sequence $(y^t, ..., y^t, z^t)_{t=0}^{\infty}$ with $(y^t, ..., y^t, z^t) \in E(\lambda^t)$, $z^t > c$ for all t. To see this, consider the following. If $z^t = a$ then from (BC) $y^t > a$ which, from (AM), implies that $\pi_1(y^t, ..., y^t, z^t) > \pi_n(y^t, ..., y^t, z^t)$, violating Proposition 1. Hence $z^t \in (a, b]$. If

$z^t = b$ then the claim is trivially true. If $z^t \in (a, b)$, then the following necessary condition for equilibrium must hold.

$$T(z^t, (n-1)y^t + z^t) + \beta_m (n-1) H_2(y^t, (n-1)y^t + z^t) = 0.$$

Since $\beta_m < 0$ when $\lambda > 1$ and $H_2 < 0$ from (NS) , we have $T(z^t, (n-1)y^t + z^t) < 0$. This, together with the fact that $z^t > y^t$ (from (AM) and Proposition 1), implies that if $z^t \leq c$, then $(n-1)y^t + z^t < nc$. But this contradicts $T(z^t, (n-1)y^t + z^t) < 0 = T(c, nc)$ since T is strictly decreasing in both components from (A1) and (SS). This proves $z^t > c$ for all t.

Since $(y^t, ..., y^t, z^t)_{t=0}^{\infty}$ converges to $(c, ..., c)$, and $z^t > c$ for all t, and $y^t < z^t$, there exists $\bar{t}$ such that for all $t > \bar{t}$, $z^t \in (c, c + \bar{\varepsilon})$ and $y^t \in (a, c + \bar{\varepsilon})$, where $\bar{\varepsilon}$ is as defined above. Hence, for all $\lambda < \lambda^t$, $\pi_n (y^t, ..., y^t, z^t) = \varphi(z^t) > \varphi(c) = \pi_i (c, ..., c)$ for all $i \in I$. In this case $\mu_r(n-1) > \mu_m(n)$ and the state $p = 1$ is unstable. ■

The above result shows that reciprocal preferences can invade a population of materialists even under (nonassortative) random matching provided that λ exceeds 1 but is not too high. The interpretation of this condition is that a single reciprocator in a group of materialists should not act too spitefully. If the degree of spite were too high, then despite the fact that the reciprocator would outperform the materialists in her group, the resulting efficiency losses would be so great as to cause her payoffs to fall below those that materialists obtain in monomorphic groups. Since most materialists will find themselves in such groups when p is close to 1, the payoff to materialists in the population as a whole will exceed that to reciprocators. Note that there is no such constraint in the case of individual selection: reciprocators can invade a group of materialists even if, in doing so, they act so spitefully as to drive their own payoffs below that which materialists earn against each other. Hence the range of parameters for which a monomorphic population of materialists is unstable is smaller under random matching than under individual selection.

If, instead of assuming common knowledge of the distribution of preferences, one assumed that individuals were completely ignorant of the preference distribution within their groups but were perfectly informed of global population composition (which then serves as a common prior in the resulting Bayesian game) then a monomorphic population of materialists could not be unstable under random matching (see Ok and Vega-Redondo, 1999, for a general analysis of this scenario). However, if there is sufficient, although not perfect, information about the preferences of players within a group, then reciprocators will be able to invade a population of materialists. For expositional clarity we demonstrate this for the case of pairwise random matching ($n = 2$). Suppose, as in the discussion of incomplete information in the previous section, that both players begin with a common prior over the distribution of preferences in their group. As in Ok and Vega-Redondo (1999), let the prior probability that any given player is a materialist be given by the global population composition $p \in (0,1)$. (In this case the prior is identical to the objective probability that any given player is a materialist under random matching.) Each player then becomes completely informed of her own preferences, and the players receive independent signals regarding the preferences of their opponents. There are two possible signals, a signal that is highly correlated with the opponent being a materialist, and one that is highly correlated with the opponent being a reciprocator. Let ρ_m be the probability that a player receives a "materialist signal", conditional on the fact that the opponent is indeed a materialist. Then $(1-\rho_m)$ is the probability that a player receives a "reciprocator signal" when her opponent is a materialist. Let ρ_r be analogously defined as the probability that a player receives a reciprocator signal, conditional on the fact that the opponent is indeed a reciprocator. This defines a Bayesian game in which there are four types of each player, where types differ not only with respect to their preferences, but also with respect to the information they receive regarding the preferences of their opponent. Let $x_{\theta i}$ denote the equilibrium action of type θ of player i, where $i \in \{1,2\}$ and $\theta \in \Theta = \{mm, mr, rm, rr\}$. Here $\theta = mm$ is a type whose preferences are materialist, and who receives a signal that her opponent is a materialist. The other types are interpreted analogously. Let $q_\theta^{\theta'}$ be the posterior probability that a type θ places on her opponent being of type θ'. It is easily verified by a straightforward application of Bayes' rule that $q_\theta^{\theta'}$ is a continuous function of ρ_m and ρ_r for any given $p \in (0,1)$.[7] The expected payoff to each type of each player may then be expressed in terms of these probabilities and the equilibrium actions of each type of each player. For instance, if θ is a type with materialist preferences ($\theta \in \{mm, mr\}$), then the expected payoffs to a player i of type θ are simply

$$v_{\theta i} = \sum_{\theta' \in \Theta} q_\theta^{\theta'} H(x_{\theta i}, x_{\theta i} + x_{\theta' j}),$$

where $i \neq j$. The expected payoffs of types with reciprocator preferences are more complicated but may easily be verified to be continuous in probabilities and actions. By continuity of the payoff functions, the correspondence mapping the signal qualities (ρ_m, ρ_r) to Nash equilibria of the corresponding Bayesian games is upper hemi-continuous. As (ρ_m, ρ_r) converges to $(1,1)$, the equilibrium actions (x_{mm1}, x_{mm2}) converge to Nash equilibria of $\Gamma(2)$,

[7]For instance, if $\theta = \theta' = mm$, then $q_\theta^{\theta'}$ is the probability that player i's opponent is a materialist conditional on the fact that player i received a materialist signal, multiplied by the probability that player i's opponent received a materialist signal conditional on the fact that player i is a materialist. The latter probability is simply ρ_m. The former probability, by application of Bayes' rule, is $p\rho_m / (p\rho_m + (1-p)(1-\rho_r))$.

the actions (x_{mr1}, x_{rm2}) and (x_{rm1}, x_{mr2}) converge to Nash equilibria of $\Gamma(1)$ and the actions (x_{rr1}, x_{rr2}) converge to Nash equilibria of $\Gamma(0)$. Hence the ordering of the equilibrium payoffs to materialists in $\Gamma(2)$ and reciprocators in $\Gamma(1)$ is preserved under incomplete information when the signals are sufficiently accurate. If the global population composition p is sufficiently close to 1, this in turn implies that under the conditions of Proposition 3, $\bar{\mu}_r(p) > \bar{\mu}_m(p)$, so that $\dot{p} < 0$. Hence the basin of attraction of the state $p = 1$ can be made arbitrarily small if the signals received regarding opponent preferences are sufficiently precise. In this sense the conclusion of Proposition 3 holds if the signals received by players about others' preferences are sufficiently precise.[8]

We conclude this section with a look at the conditions under which a population of reciprocators is stable under random matching.

Proposition 4. *Consider any separable* $\Gamma \in \mathcal{A}$ *which satisfies* (AM), (NS), (A1), (SS), *and* (BC). *There exists* $\tilde{\lambda} < n-1$ *such that if* $\lambda > \tilde{\lambda}$, *a monomorphic population of reciprocators is stable.*

Proof.

Claim 1. If $\Gamma \in \mathcal{A}$ is separable and satisfies (A1) , and (SS), then $\Gamma(0)$ has a unique equilibrium.

Proof of Claim 1. From Lemma 1, all equilibria of $\Gamma(0)$ are symmetric. Let $(d, ..., d)$ and $(d', ..., d')$ be two equilibria with $d > d'$. Then the following are necessary equilibrium conditions.

$$\begin{aligned} T(d, nd) + \beta_r(n-1)H_2(d, nd) &\geq 0, \\ T(d', nd') + \beta_r(n-1)H_2(d', nd') &\leq 0. \end{aligned}$$

Since T is decreasing in both components, $T(d, nd) < T(d', nd')$, so the above conditions imply that $H_2(d', nd') < H_2(d, nd)$. Since $H_{12} = 0$, $H_2(d', nd') = H_2(d, nd')$ so we have $H_2(d, nd') < H_2(d, nd)$. But when $H_{12} = 0$, (SS) implies that $H_{22} < 0$ and hence $H_2(d, nd') > H_2(d, nd)$, a contradiction. ||

Claim 2. Suppose $\Gamma \in \mathcal{A}$ satisfies (AM), (A1) , (SS) and $H_{12} \geq 0$, and that $\lambda = n - 1$. Then $\Gamma(1)$ and $\Gamma(n)$ have the same (unique) equilibrium.

[8] If the global population if finite, even a single mutation will cause p to be bounded away from 1. In this case it can be proved, using the above reasoning, that the state $p = 1$ is unstable. For an infinite population, the instability of the state $p = 1$ does not follow because for any given values of signal quality $(\rho_m, \rho_r) \in (0, 1)^2$ it is possible to find a number $\bar{p}$ sufficiently close to 1 such that if the prior $p > \bar{p}$, then the posterior probability that one is facing a materialist can be close to 1 regardless of the signal received.

Proof of Claim 2. For a proof that $\Gamma(n)$ has a unique equilibrium under (A1) and (SS), see Corchón (1996, Proposition 1.3). Let $(c, ..., c)$ denote this equilibrium. From Lemma 1, any equilibrium x of $\Gamma(1)$ is of the form $(y, z, ..., z)$. Since $\lambda = n - 1$, $(\beta_r(n-2) + \beta_m) = 0$ so for all $i \in R$,

$$\frac{\partial u_i}{\partial x_i} = T(z, n\bar{x}) + (\beta_r(n-2) + \beta_m)H_2(y, n\bar{x}) = T(z, n\bar{x}).$$

If $y < z$, then equilibrium requires that $T(y, n\bar{x}) \leq 0 \leq T(z, n\bar{x})$ which from (A1) implies that $y \geq z$, a contradiction. Similarly, if $y > z$, then $T(z, n\bar{x}) \leq 0 \leq T(y, n\bar{x})$ which from (A1) implies that $z \geq y$, a contradiction. Hence $y = z = \bar{x}$ and $T(y, ny) = T(z, nz) = 0$. But since $T(c, nc) = 0$, $y = z = c$ from (A1). ||

From Claim 1, $\Gamma(0)$ has a unique equilibrium, which is therefore symmetric and which we denote by $(d, ..., d)$. From Claim 2, if $\lambda = n - 1$, $\Gamma(1)$ has the same unique equilibrium as $\Gamma(n)$, which we denote by $(c, ..., c)$. From Proposition 5 below, there exists $\varepsilon > 0$ such that $\pi_i(d, ..., d) = \pi_i(c, ..., c) + \varepsilon$ for all $i \in I$. Hence, when $\lambda = n - 1$, $\mu_m(1) < \mu_r(0)$ so the equilibrium at $p = 0$ is stable.

Next we show that there exists $\tilde{\lambda} < n - 1$ such that the result holds for $\tilde{\lambda} < \lambda < n - 1$. With $\alpha > 0$ given, let $E(\lambda)$ denote the set of equilibria of $\Gamma(1)$ when the reciprocator has preference parameter λ. Consider a sequence $\left(\lambda^t\right)_{t=0}^{\infty}$ where $\lambda^{t-1} < \lambda^t < n - 1$ for all $t \geq 1$ and $\lim_{t\to\infty} \lambda^t = n - 1$. Since the Nash equilibrium correspondence has a closed graph, any sequence $(x^t)_{t=0}^{\infty}$ with $x^t \in E(\lambda^t)$ has a limit point in $E(n-1)$. Since $E(n-1)$ consists of the single element $(c, ..., c)$ and any sequence $(x^t)_{t=0}^{\infty}$ with $x^t \in E(\lambda^t)$ converges to $(c, ..., c)$. Since the payoff functions are continuous, there exists $\tilde{t}$ such that for all $t > \tilde{t}$ and all $i \in I$, $\pi_i(x^t) < \pi_i(c, ..., c) + \varepsilon$, where ε is as defined above. Setting $\tilde{\lambda} = \lambda^{\tilde{t}}$, we have the following: if $\tilde{\lambda} < \lambda < n-1$, then for any equilibrium x of $\Gamma(1)$, $\pi_i(x) < \pi_i(d, ..., d)$ for all $i \in M$. Hence $\mu_m(1) < \mu_r(0)$ when $\tilde{\lambda} < \lambda < n - 1$, so the equilibrium at $p = 0$ is stable.

To complete the proof, consider the case $\lambda > n - 1$. From Lemma 1, any equilibrium x of $\Gamma(1)$ is of the form $x = (y, z, ..., z)$. From Proposition 2 and (AM), $y < z$, so $T(y, n\bar{x}) \leq 0$ is a necessary equilibrium condition. We claim that $c < \bar{x}$. To see this, suppose $\bar{x} \leq c$. Then $y < c$ (otherwise we would have $c \leq y < z$ contradicting $\bar{x} \leq c$), which implies that $T(y, n\bar{x}) > 0 = T(c, nc)$ from (A1) and (SS), contradicting $T(y, n\bar{x}) \leq 0$. Hence $\bar{x} > c$. We next claim that $y < c$. To see this, suppose $c \leq y$. Then, since $c < \bar{x}$, (A1) and (SS) imply $T(y, n\bar{x}) < 0 = T(c, nc)$. But $T(y, n\bar{x}) < 0$ can hold in equilibrium only if $y = a < c$, contradicting $c \leq y$. We have therefore proved that $y < c < \bar{x}$. This, together with (AM) and (NS), implies $H(y, n\bar{x}) < H(c, n\bar{x}) < H(c, nc)$. But $H(c, nc) < H(d, nd)$ from Proposition 5 below. Hence $\mu_m(1) = H(y, n\bar{x}) < H(d, nd) = \mu_r(0)$ when $\lambda > n - 1$, so the equilibrium at $p = 0$ is stable. ■

Proposition 4 confirms that reciprocal preferences can persist in competition with materialist preferences under (nonassortative) random matching, and that it is possible in this environment for materialist preferences to be eliminated entirely. Furthermore, in contrast with the case of individual selection, a population of reciprocators can resist invasion by materialists even when the presence of a single materialist in a group of reciprocators does not cause the latter to become spiteful. This follows from the fact that the threshold $\tilde{\lambda} < n - 1$ in Proposition 4. The intuition for this is as follows. When a single materialist is present in a group of reciprocators, the latter continue to remain altruistic but become less so. Provided that the reduction in altruism is sufficiently great ($\lambda > \tilde{\lambda}$) the average group payoff is lowered significantly relative to the case of groups containing only reciprocators. Hence, although the single materialist outperforms the reciprocators in her group, her payoff is lower than that which reciprocators earn in monomorphic groups. Since almost all reciprocators find themselves in monomorphic groups when p is sufficiently small, materialists cannot invade.

We next turn to the question of efficiency, which is critical in understanding whether reciprocal preferences are favored under assortative interaction and group selection.

6 Efficiency

There are at least two reasons why the issue of efficiency is important for understanding the evolution of preferences. First, if group selection is an important force in determining the fate of populations, for instance through the collapse or extinction of poorly performing groups, then preferences that are efficiency enhancing are liable to be favored. Second, if group formation occurs under voluntary association rather than random matching, then it may be advantageous for those with efficiency enhancing preferences to seek each other out in the formation of groups. Most evolutionary explanations of pure altruism are based on one or both of these processes of group selection and assortative interaction (Sober and Wilson, 1998). Pure altruism, however, suffers from evolutionary disadvantages under individual selection or random matching in many strategic environments. In contrast, reciprocal preferences can yield some of the same group benefits that altruism does, without being vulnerable in competition with materialist preferences under individual selection. The following result is a formal statement of the fact that groups of reciprocators outperform groups of materialists.

Proposition 5. *Suppose* $\Gamma \in \mathcal{A}$ *is separable satisfies* (NS), (A1), (SS), *and* (BC). *Then, if* x *is an equilibrium of* $\Gamma(0)$ *and* y *is an equilibrium of* $\Gamma(n)$, $\pi_i(x) > \pi_i(y)$ *for all* $i \in I$.

Proof. Consider any symmetric action profile $x = (z, ..., z)$ where $z \in [a, b]$. The payoff to each player at x is given by $W(z) = H(z, nz)$. Note that $W'(z) = H_1(z, nz) + nH_2(z, nz) = T(z, nz) + (n-1)H_2(z, nz)$. Conditions (A1), (SS) and $H_{12} = 0$ together imply that

$$W''(z) = H_{11}(z, nz) + 2nH_{12}(z, nz) + n^2 H_{22}(z, nz) < 0. \tag{5}$$

Let $e = \text{argmax}_{z\in[a,b]} W(z)$. This is the action which, if taken by all players, yields the highest payoff to each among the set of symmetric action profiles.

Let x^m be an equilibrium of $\Gamma(n)$ and x^r an equilibrium of $\Gamma(0)$. From Lemma 1, equilibria of $\Gamma(n)$ and $\Gamma(0)$ are symmetric under the stated conditions. Hence there exist $c, d \in [a, b]$ such that $x^m = (c, ..., c)$ and $x^r = (d, ..., d)$. From (BC), $c \in (a, b)$. We claim that $d \in [a, b)$. To see why, note that for $x = (b, ..., b)$ to be an equilibrium of $\Gamma(0)$,we must have $\partial u_i/\partial x_i \geq 0$ for all $i \in I$. But at $x = (b, ..., b)$, (BC) implies that $T(b, nb) < 0$ and from (NS) we therefore have

$$\frac{\partial u_i}{\partial x_i} = T(b, nb) + \beta_r(n-1)H_2(b, nb) < 0,$$

a contradiction. Hence $d \in [a, b)$. Consider the following two cases.

(i) Suppose $d = a$. Then a necessary condition for equilibrium is

$$\frac{\partial u_i}{\partial x_i} = T(a, na) + \beta_r (n-1) H_2(a, na) \leq 0.$$

But since $\beta_r \in (0, 1)$ and $H_2 < 0$ from (NS), this implies that

$$T(a, na) + (n-1) H_2(a, na) = W'(a, na) < 0.$$

The above, together with (5), implies that $e = d = a$. Since $c > a$, and $e = \text{argmax}_{z\in[a,b]} W(z)$, we have $W(e) = W(d) > W(c)$ as required.

(ii) Suppose $d \in (a, b)$. Consider the following function

$$G(z, \beta) = H_1(z, nz) + H_2(z, nz) + \beta (n-1) H_2(z, nz).$$

Note that if $\beta = 0$, $G(z, \beta) = 0$ is a necessary condition for equilibrium in $\Gamma(n)$; if $\beta = \beta_r$, $G(z, \beta) = 0$ is a necessary condition for equilibrium in $\Gamma(0)$, and if $\beta = 1$, $G(z, \beta) = 0$ corresponds to the condition $W'(z) = 0$. Note also that

$$\frac{\partial G}{\partial z} = H_{11} + nH_{12} + (1 + \beta (n-1)) (H_{12} + nH_{22}) < 0$$

for all $\beta \in [0,1]$ from (A1) , (SS)and $H_{12} = 0$. Applying the implicit function theorem, $G(z, \beta) = 0$ defines a function $z(\beta) : [0,1] \to \mathbf{R}$ with the property

$$\frac{dz}{d\beta} = -\frac{\partial G/\partial \beta}{\partial G/\partial z} < 0$$

since $\partial G/\partial \beta = (n-1) H_2(z, nz) < 0$ from (NS). Hence $z(1) < z(\beta_r) < z(0)$. If $z(1) < a$ then $e = a$; otherwise $e = z(1)$. In either case, $e < z(\beta_r) = d < z(0) = c$, so from (5) and the fact that $e = \text{argmax}_{z \in [a,b]} W(z)$, we have $W(d) > W(c)$ as required. ■

The above result suggests that group selection or perfectly assortative interaction should favor the growth of reciprocators over groups of materialists. However, although a monomorphic group of reciprocators does better than a monomorphic group of materialists, it is not the case that the average payoff in a group increases monotonically with the number of reciprocators. In groups consisting largely of materialists, reciprocators act in a spiteful manner, choosing higher actions in equilibrium than would be optimal from a purely material standpoint. This can cause groups with a small number of reciprocators to obtain lower average payoffs than monomorphic groups of either type. The following numerical example illustrates this.

Example 3. Suppose Γ is a common pool resource game (see Example 2) game with $A(X) = 10 - X$, $w = 1$, $n = 20$, $\lambda = 2$, and $\alpha = 0.5$. It can be shown that equilibria of $\Gamma(k)$ are unique for all k. Let $\bar{\pi}(k)$ be the mean equilibrium payoff in the group when the population composition is k. Computation of equilibria yields $\bar{\pi}(n) = 0.184 < 0.578 = \bar{\pi}(0)$. However, $\bar{\pi}(k)$ does not decline monotonically with k, as shown in Figure 1.

The fact that reciprocators can be efficiency-reducing when they are rare suggests that under group selection, mixed groups will tend to have the lowest prospects for survival. The groups which proliferate fastest will be monomorphic groups of reciprocators, which (from Proposition 5) outperform monomorphic groups of materialists. Although we do not explore the effects of group selection formally in the present paper, it is easy to construct models of intergroup competition in which the efficiency-enhancing effects of reciprocal preferences (when they are sufficiently widespread) causes such preferences to outcompete purely self-regarding preferences. Group selection in general tends to favor the survival of efficiency-enhancing traits (see, for instance, Canals and Vega-Redondo, 1998, and the references cited therein.)

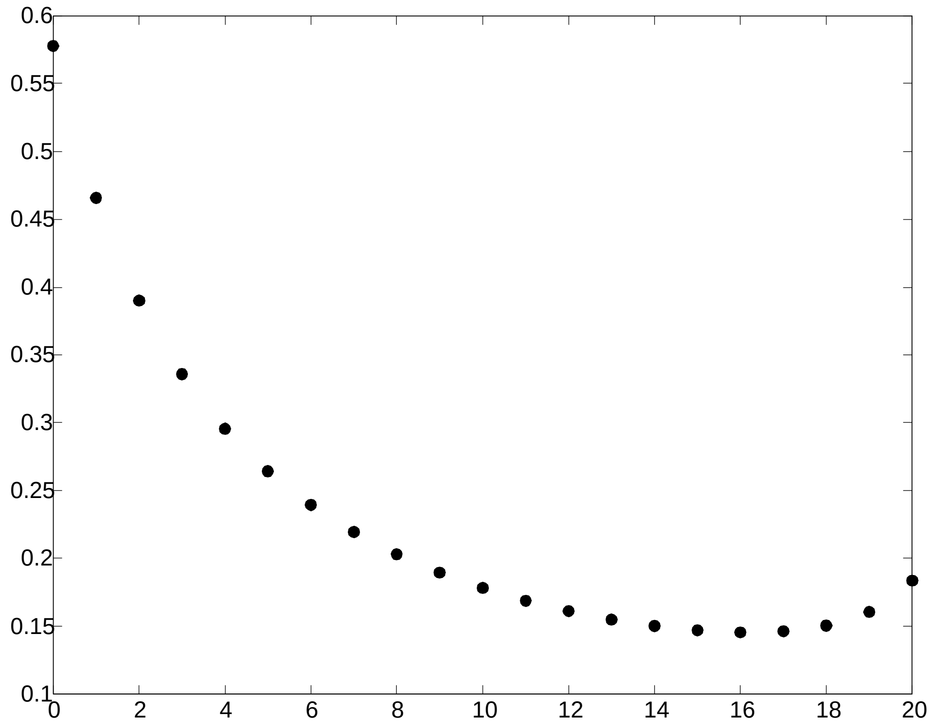

Figure 1. Average group payoffs as a function of population composition.

Finally, consider the case of assortative interaction in which individuals form groups by voluntary association. This would result in perfect assortation regardless of λ. To see this, note that if $\lambda > n-1$ then materialists would prefer to associate exclusively with each other, since even a single materialist in the presence of reciprocators would cause the latter to become spiteful. If, on the other hand, $\lambda < n-1$, then reciprocators will associate exclusively with each other. This is because the presence of a single materialist in a group of reciprocators both lowers the average payoff in the group and results in a higher payoff for the materialist relative to the reciprocators. These two facts together imply that the material payoff of each reciprocator is strictly lowered. Given that their objective function places positive weight on the payoffs of other reciprocators and negative weight on the payoff of materialist, this implies a lower value of their objective function. Consequently, reciprocators will prefer to associate exclusively with each other, leading to perfect assortation. Proposition 5 then implies that reciprocators will outperform materialists in the population as a whole.

7 Conclusions

The analysis in this paper suggests that a population of self-regarding materialists may be unstable in the presence of reciprocal preferences under a variety of evolutionary selection processes, including individual selection, random matching, group selection and assortative interaction. Individuals endowed with such preferences are willing to make material sacrifices to reward others who are similarly disposed, and to punish those who are not. Their motivation for doing so does not arise from any prospects of future material reward. Such preferences not only help account for experimental data from a diverse set of sources, they also accord with the facts of everyday experience. Even without any history of prior interaction, and with little or no prospect of future interaction, people are often altruistic towards others who are perceived to be similarly altruistic, and may even gain pleasure from reducing the well being of those who are perceived to be selfish or spiteful. Such behavior has increasingly come to be recognized as an important aspect of human decision making with significant social and economic implications such as the downward rigidity of real wages, the private provision of certain public goods, the sustainable management of natural resources in local commons, voluntary donations of time and effort, and the decentralized enforcement of cooperative social norms (see Fehr and Gächter, 1998, for a recent survey of the relevant literature.)

A natural extension of the present work would be the endogenization of the preference parameters. The class of preferences considered here is large and varies along two dimensions: the degree of altruism and the degree of sensitivity to the altruism of others. While a wide range of parameter values is consistent with survival against materialists, a much narrower range may be expected to survive when several members of this class of preferences are in competition with each other. Another possible extension of this work would be to study the evolution of reciprocal preferences in other environments likely to have been important in the evolution of human behavior, such as multi-stage games which allow for the costly sanctioning of prior actions. Under incomplete information, individuals would be induced to take into account the effect of their actions on the beliefs of others regarding the distribution of preferences (as in Kreps *et al.*, 1982, for instance.) An evolutionary analysis that allows for such signalling effects could potentially yield significant new insights.

References

[1] Andreoni, J. and J.H. Miller (1999). "Giving according to GARP: An Experimental Test of the Rationality of Altruism." SSRI Working Paper 9902, University of Wisconsin.

[2] Axelrod, R. and W.D. Hamilton (1981). "The Evolution of Cooperation." *Science* 211: 1390–1396.

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

[4] Blount, S. (1995). "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes* 63: 131–44.

[5] Bolton, G. (1991): "A Comparative Model of Bargaining: Theory and Evidence." *American Economic Review* 81: 1096-1136.

[6] Bolton, G. and A. Ockenfels (1998). "ERC: A Theory of Equity, Reciprocity and Competition." Mimeo, Pennsylvania State University.

[7] Bowles, S. and H. Gintis (1998). "The Evolution of Strong Reciprocity." Santa Fe Institute Working Paper 98-08-073E.

[8] Boyd, R. and P.J. Richerson (1989). "The Evolution of Indirect Reciprocity." *Social Networks* 11: 213–236.

[9] Canals, J. and F. Vega-Redondo (1998). "Multi-level Evolution in Population Games." *International Journal of Game Theory* 27: 21–35.

[10] Corchón, L. (1996). *Theories of Imperfectly Competitive Markets.* Berlin: Springer-Verlag.

[11] Dekel, E., J. Ely and O. Yilankaya (1998). "Evolution of Preferences." Mimeo, Northwestern University.

[12] Dubey, P., Mas-Colell, A., and Shubik, M. (1980). "Efficiency Properties of Strategic Market Games." *Journal of Economic Theory* 22: 339-362.

[13] Dufwenberg, M. and G. Kirchsteiger (1998). "A Theory of Sequential Reciprocity." CentER Discussion paper 9837, Tilburg University.

[14] Ely, J. and O. Yilankaya (1997). "Evolution of Preferences and Nash Equilibrium." Mimeo, Northwestern University.

[15] Falk A. and U. Fischbacher (1998). "A Theory of Reciprocity." Mimeo, University of Zurich.

[16] Fehr, E. and S. Gächter (1998). "Reciprocity and Economics: The Economic Implications of *Homo Reciprocans." European Economic Review* 42: 845–59.

[17] Fehr, E., G. Kirchsteiger, and A. Reidl (1993). "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics* 108: 437–460.

[18] Fehr, E. and K.M. Schmidt (1997). "A Theory of Fairness, Competition, and Cooperation." Mimeo, University of Zurich.

[19] Fudenberg, D. and E.S. Maskin (1986). "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54: 533–554.

[20] Geanakopolos, J., D. Pearce and E. Stacchetti (1989), "Psychological Games and Sequential Rationality." *Games and Economic Behavior* 1: 60–79.

[21] Güth, W., R. Schmittberger and B. Schwarze (1982). "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* 3: 367–388.

[22] Güth, W., and M. Yaari (1992). "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in U. Witt, ed., *Explaining Forces and Change: Approaches to Evolutionary Economics.* Ann Arbor: University of Michigan Press.

[23] Isaac, R.M. and J.M. Walker (1988). "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism." *Quarterly Journal of Economics* 103: 179–200.

[24] Kreps, D., P. Milgrom, J. Roberts and R. Wilson (1982). "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma." *Journal of Economic Theory* 27: 245–52.

[25] Kahneman, D., J.L. Knetsch and R. Thaler (1986). "Fairness and the Assumptions of Economics." *Journal of Business* 59: S285–300.

[26] Koçkesen, L., Ok, E.A., and Sethi, R. (1997). "The Strategic Advantage of Negatively Interdependent Preferences." C.V. Starr Center Working Paper 97-34, NYU.

[27] Koçkesen, L., Ok, E.A., and Sethi, R. (1998). "Evolution of Interdependent Preferences in Aggregative Games." *Games and Economic Behavior*, forthcoming.

[28] Levine, D. (1998). "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1: 593–622.

[29] Nowak, M.A. and K. Sigmund (1998). "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393: 573–577.

[30] Ok, E.A. and F. Vega-Redondo (1999). "On the Evolution of Individualistic Preferences: Complete Versus Incomplete Information Scenarios." Mimeo, NYU.

[31] Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83: 1281–1302.

[32] Roth, A.E., V. Prasniker, M. Okuno-Fujiwara, and S. Zamir (1991). "Bargaining and Market Behavior in Jerusalem, Liubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review* 81: 1068–1095.

[33] Smith, V. (1982). "Microeconomic Systems and an Experimental Science." *American Economic Review* 72: 923–955.

[34] Sober, E. and D.S. Wilson (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior.* Harvard University Press.

[35] Trivers, R.L. (1971). "The Evolution of Reciprocal Altruism" *Quarterly Review of Biology* 46: 35–57.