Journal of Economic Theory **97**, 273–297 (2001)
doi:10.1006/jeth.2000.2683, available online at http://www.idealibrary.com on IDEAL®

Preference Evolution and Reciprocity[1]

Rajiv Sethi

Department of Economics, Barnard College, Columbia University, 3009 Broadway, New York, New York 10027
rs328@columbia.edu

and

E. Somanathan

Department of Economics, University of Michigan, 611 Tappan Street, Ann Arbor, Michigan 48109
esomana@umich.edu

Received April 30, 1999; final version received April 14, 2000;
published online March 6, 2001

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 assortative and nonassortative matching. In comparison with simpler specifications of preference interdependence (such as pure altruism or spite), the survival of such preferences is therefore less sensitive to details of the evolutionary selection process. *Journal of Economic Literature* Classification Numbers: C72, D62.

Key Words: reciprocity; evolution; preference interdependence.

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 [35], Roth *et al.* [34]), but fail rather dramatically in others, such as public goods, ultimatum bargaining, and gift

[1] We thank Levent Koçkesen, Efe Ok, two anonymous referees, and especially Larry Samuelson for many suggestions that have improved this paper. Conference participants at Penn State University and the University of Massachusetts and seminar participants at McGill University also provided helpful comments. The first author gratefully acknowledges financial support from the National Science Foundation under Grant SBR-9812379.

0022-0531/01 $35.00

exchange games (Isaac and Walker [25], Güth *et al.* [22], Fehr *et al.* [17]). 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 [1], Bolton [4]). 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 [33], Fehr and Schmidt [18], Bolton and Ockenfels [6], Levine [31], Falk and Fischbacher [15], and Dufwenberg and Kirchsteiger [12].) 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 such 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* similar to those which Levine [31] 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. We argue that the flexibility in behavior that reciprocal preferences provide enables such preferences to survive under evolutionary pressure within a class of environments under both assortative and purely random (nonassortative) interaction. Under assortative interaction, reciprocators are likely to find themselves in groups consisting largely of other reciprocators. This leads them to behave altruistically and enjoy the efficiency gains that altruism provides in many strategic situations. When interaction is purely random or nonassortative, and the global population consists predominantly of materialists, reciprocators are likely to find themselves in groups consisting largely of materialists. This leads them to act as if they had spiteful preferences and, under certain conditions, induces a response from materialists that raises

the material payoffs of reciprocators. If the spitefulness of reciprocators toward materialists is not too great, this effect can lead to a higher expected payoff for reciprocators relative to that of materialists in the global population. This, in turn, permits preferences for reciprocity to survive and spread when the population share of materialists is sufficiently large. Once the population share of reciprocators becomes large, on the other hand, most groups will consist largely of reciprocators. For parameter values within a certain range, materialists can thrive in such groups since 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. In this case a polymorphic population (consisting of both materialists and reciprocators) will prevail in evolutionary equilibrium.

The model we propose has the following features. There is a large population of individuals who are matched in small subgroups in which they interact strategically. 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. While individuals act to maximize their utility (which may depend on the material payoffs and preferences of others), it is their realized material payoffs that determine the evolutionary survival of their preferences. We consider first the case of non-assortative matching: at the end of each period of interaction, all individuals are randomly matched with others in the global population to form new groups. It is shown that a population of materialists will not generally be stable in the presence of reciprocators. We next 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. 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.

It is assumed that the strategic interaction that occurs within groups belongs to the class 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.* [11], Corchón [9]), 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. While our focus on aggregative games is motivated primarily by their analytical tractability, we consider them to be a reasonable class of strategic environments within which to study the problem of preference evolution.

Before proceeding, it is useful to compare our approach to the evolution of preferences with earlier (and ongoing) work on the topic. The literature on preference evolution in games was pioneered by Güth and Yaari [23], and has been further developed by Güth [21], Bowles and Gintis [7], and Huck and Oechssler [24] among others. In each of these papers, the definition of reciprocal preferences is itself tailored to the specific environment under consideration. For instance, Güth and Yaari allow for individuals who have a preference for rejecting unfair offers in bargaining games, while Bowles and Gintis consider individuals with a taste for punishing free riders in a model of team production. In contrast, Bester and Güth [3] and Koçkesen *et al.* [29–30], have considered more general specifications of preferences that are defined independently of particular strategic environments, and depend only on the distribution of material payoffs in the group. Bester and Güth deal with the survival of altruistic preferences under pairwise random matching and Kçkesen *et al.* with the survival of envious or spiteful preferences. Ely and Yilankaya [13] and Dekel *et al.* [10] examine general models of preference evolution in which the class of preferences is composed of all possible orderings over action profiles. Reciprocal preferences of the kind considered in the present paper are not encompassed by this class of preferences for the reason that, in our case, individual orderings over action profiles are sensitive to the preferences of other players.[2]

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 π_i: $\times_j X_j \to \mathbf{R}$, $i \in I$, the *material* payoff functions. In the context of experimental games, material payoffs correspond to cash payments. More generally, material

[2] There is also a large literature on the evolution of reciprocity as an aspect of *behavior* rather than an attribute of preferences. Early contributions include Trivers [37] and Axelrod and Hamilton [2], who showed that direct reciprocity (interpreted as a Tit-for-Tat behavioral strategy) can be sustained under evolutionary competition when interactions are repeated indefinitely. Such behavior is known to be consistent 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 [19]).

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 his of her 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: $\times_j X_j \to \mathbf{R}$ 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 [18] and Bolton and Ockenfels [6] 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 for the fact that at least in some environments, subjects consistently choose very different terminal payoff distributions depending on the prior behavior and opportunities available to other players.[3]

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 [33], Dufwenberg and Kirchsteiger [12], Falk and Fischbacher, [15]).[4] In equilibrium, all beliefs (including higher-order

[3] For instance, identical offers in a ultimatum bargaining games "trigger vastly different rejection rates depending on the other offers available to the proposer" (Falk *et al.* [14]). See Kagel *et al.* [26] for additional arguments along these lines and Bolton *et al.* [5] for a dissenting view.

[4] See Geanakoplos *et al.* [20] for the methodology and fundamental properties of psychological games.

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 him or her. These papers are effective in accounting for the role of intentionality in experimental results. As presently formulated, however, they deal only with two-person games and therefore cannot be used to address phenomena such as the rewarding of individuals who have been kind to *others*.[5]

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. Levine [31] suggests the following specification of reciprocal preferences, which allow for both 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 $0 \leqslant \lambda_i$. 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

[5] As an example, consider the results of Kahneman *et al.* [28], 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.

under nonassortative matching in the class of strategic environments considered here. The following slight variant of Levine's specification, however, has 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 $0 \leqslant \alpha_i < 1$ and $\lambda_i \geqslant 0$. 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. Note that materialist preferences may be viewed as a limit case of reciprocator preferences, obtained as α and λ approach zero.

The preferences of reciprocators possess the property that an individual i's utility depends not only on another individual j's payoff, but the extent of this dependence itself depends on j's preferences. It is this property that allows individuals the flexibility to switch from altruistic to spiteful behavior in response to the composition of the group within which they find themselves. The capacity for spite is necessary for the viability of such preferences under random matching, and the capacity for altruism is a critical ingredient for their viability under assortative matching. One could obtain some of the results below by assuming preferences that were purely spiteful and others by assuming preferences that were purely altruistic. For instance, altruistic preferences can be viable under perfectly assortative interaction but not under random matching, and spiteful preferences can be viable under random but not assortative interaction.[6] Reciprocal preferences, on the other hand, can be viable in evolutionary competition with materialist preferences under both assortative and nonassortative matching. We demonstrate this within a particular class of strategic environments: aggregative games.

[6] Specifically, Propositions 1 and 3 below would remain valid if, instead of reciprocators, we considered purely spiteful players. (In the case of the latter result the level of spite cannot be too great.) Proposition 5 would remain valid if, instead of reciprocators, we considered purely altruistic players.

3. AGGREGATIVE GAMES

Although the analysis of preference evolution can, in principle, be conducted within arbitrary strategic environments, with a population of possible preferences for each player position, it is reasonable to begin with the case of a single population from whom individuals are drawn to play a game that is symmetric with respect to material payoffs. In this paper attention is confined to games $\Gamma \equiv \{X, \pi_i\}_{i \in I}$ that are symmetric in this sense. Here 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.) We further 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.* [11]). 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:

$$H_1 > 0 \tag{4}$$

$$H_2 < 0 \tag{5}$$

$$H_{11} + H_{21} = T_1 < 0 \tag{6}$$

$$H_{21} + H_{22} = T_2 > 0. \tag{7}$$

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.[7] Assumptions (6) and (7) 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 (6) and (7) together

[7] 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 (4–5) continue to hold if the signs of *both* inequalities are reversed.

imply strict concavity of payoffs in own actions. Both these assumptions are common in analyses of aggregative games (see, for instance, Corchón [9]), together with the following conditions, which are made to exclude boundary equilibria of limited interest in the present context.

$$T(a, na) > 0 > T(b, nb). \tag{8}$$

Let $\mathscr{A}$ denote the class of symmetric aggregative games which satisfy conditions (4–8). The following examples illustrate that this class includes games which have economically meaningful interpretations and which are relevant environments in which the question of preference evolution may be examined.

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 (4–7). If, in addition, $f'(0) > g'(nb)$ and $f'(nb) < g'(0)$, it satisfies (8).

EXAMPLE 2 (Common pool resource extraction). Suppose each of n individuals has access to a common pool resource. Let $x_i \geqslant 0$ 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 (4–8).[8]

As a special case to be considered below, we say that a game $\Gamma \in \mathscr{A}$ 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

[8] Since $A' < 0$, the assumption $nA' + n\bar{x}A'' < 0$ is trivially satisfied if $A'' \leqslant 0$. 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 sometimes assumed in this context.

described in Example 1 is separable, but that in Example 2 is not. The separability condition may be interpreted, in the context of the public goods game, as an assumption that the value to an individual of private good consumption does not depend on the amount of public good provided.

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 for all $i \in R$, $u_i(x)$ is given by (1–2) for some value of $\alpha \in (0, 1)$ and $\lambda > 0$. 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

$$u_i(x) = \begin{cases} \pi_i(x) & \text{for all} \quad i \in M, \\ \pi_i(x) + \beta_r \sum_{j \in R \setminus \{i\}} \pi_j(x) + \beta_m \sum_{j \in M} \pi_j(x) & \text{for all} \quad i \in R, \end{cases} \tag{9}$$

where β_r and β_m satisfy

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

Let $\Gamma(k)$ denote this game, where $k \in \{0, ..., n\}$ is the number of players with materialist preferences. 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. 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.

We conclude this section with some preliminary results which will prove useful in the subsequent analysis. The first of these establishes conditions under which a single reciprocator in a group of materialists obtains greater payoffs than each of the materialists.

PROPOSITION 1. *Suppose $\Gamma \in \mathscr{A}$. Then, at any equilibrium of $\Gamma(n-1)$, the single reciprocator earns a strictly greater (smaller) payoff than each materialist if $\lambda > 1$ ($\lambda < 1$).*

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 $T(x_i, n\bar{x}) > T(x_j, n\bar{x})$. This implies $x_i \leqslant x_j$ from (6), a contradiction. Hence $x = (y, ..., y, z)$ for some $y, z \in [a, b]$. If $z = b$ or $y = a$, then $z > y$ from (8), and so $\pi_n > \pi_i$ for all $i \in M$ from (4). Now suppose $z < b$ and $y > a$. In this case $T(y, n\bar{x}) \geqslant 0$ and

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

Since $H_2 < 0$ from (5) and $\beta_m < 0$, this implies $T(z, n\bar{x}) < 0 \leqslant T(y, n\bar{x})$. Hence $z > y$ from (6), and so $\pi_n > \pi_i$ for all $i \in M$ from (4). A similar argument can be used to show that if $\lambda < 1$, then $\pi_n < \pi_i$ for all $i \in M$. (Note that (7) is not required for the result to hold.) ▌

The requirement that $\lambda > 1$ for the above to hold is intuitive, since reciprocators are altruistic even towards materialists when $\lambda < 1$. It is the potentially spiteful behavior of the single reciprocator which gives him or her the advantage over materialists.

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 $x_i = x_j$ whenever 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* Γ *satisfies* (6) *and* $H_{12} \geqslant 0$. *Then, for any* $k \in \{0, ..., n\}$, *every equilibrium 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 $\partial u_i/\partial x_i \geqslant 0 \geqslant \partial u_j/\partial x_j$, or

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

where $\beta_r > 0$. This implies $T(x_i, n\bar{x}) + \beta_r H_2(x_j, n\bar{x}) \geqslant T(x_j, n\bar{x}) + \beta_r H_2(x_i, n\bar{x})$. Since $H_{12} > 0$, $H_2(x_i, n\bar{x}) \geqslant H_2(x_j, n\bar{x})$. Hence $T(x_i, n\bar{x}) \geqslant T(x_j, n\bar{x})$. This implies $x_i \leqslant x_j$ from (6), 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 $H_{12} \geqslant 0$ is not required in this latter case.) ▌

The following result identifies, in the special case of separable payoff functions, the relevant parameter range for which a single materialist in a group of reciprocators obtains greater payoffs than each of the reciprocators.

PROPOSITION 2. *Consider any separable* $\Gamma \in \mathscr{A}$. *Then, at any equilibrium of* $\Gamma(1)$, *the single materialist earns a strictly greater* (*smaller*) *payoff than each reciprocator if* $\lambda < n-1$ $(\lambda > n-1)$.

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 (8), and so $\pi_1 > \pi_i$, for all $i \in R$ from (4). Now suppose $y < b$ and $z > a$. Then $T(y, n\bar{x}) \leqslant 0$ and

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

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

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

Since $\lambda < n-1$, $\beta_r(n-2) + \beta_m = (n-1-\lambda)\,\alpha/(1+\lambda) > 0$. This, together with (5) and the above relation yields $T(z, n\bar{x}) > 0 \geqslant T(y, n\bar{x})$. Hence $z < y$ from (6), and $\pi_1 > \pi_i$ for all $i \in R$ from (4). A similar argument can be used to show that if $\lambda > n-1$, then $\pi_1 < \pi_i$ for all $i \in R$. ∎

Taken together, Propositions 1 and 2 imply that when $1 < \lambda < n-1$, a single materialist in a group of reciprocators will outperform all reciprocators in that group, while a single reciprocator in a group of materialists will outperform all materialists in that group. The intuition underlying this is the following. In a group 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 group 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.[9] 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. For given λ, Proposition 2 implies that if the group size n is sufficiently large, then a single materialist in a population of reciprocators will outperform all others in the group, as their altruism for each other restrains them from responding to the presence of the materialist in a spiteful manner. In other words, unless the group size is sufficiently small, a single materialist will thrive in a group of reciprocators. This has evolutionary implications that are discussed in the section to follow.

4. RANDOM MATCHING

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 of size n. 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 equilibrium payoff to materialists in groups with population composition k, and let $\mu_r(k)$ be the expected equilibrium payoff to reciprocators. When all players in a group are materialists the resulting game $\Gamma(n)$ has a unique equilibrium, and when there is a single reciprocator in a group the resulting game $\Gamma(n-1)$ has a unique

[9] Note that this effect occurs because reciprocators cannot reduce the payoffs of materialists without simultaneously lowering those of fellow-reciprocators. If reciprocators could target materialists individually (for instance, if the game were augmented to allow a second stage in which costly sanctions could be imposed on specific individuals), a single materialist in a group of reciprocators might do very poorly in comparison with the reciprocators even when λ is close to 1. A proper analysis of games with this structure, while interesting, is beyond the scope of the present paper.

equilibrium provided that λ is sufficiently close to 1. More generally, however, there may be multiple equilibria and the payoffs $\mu_m(k)$ and $\mu_r(k)$ will then 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 $\Gamma \in \mathscr{A}$. 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 (6) and (7), $G_1 = T_1 + (n-1)\, T_2 < 0$ 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 (6) and (7), $\Gamma(n)$ has a unique equilibrium (Corchón [9, Proposition 1.3]), which must therefore be symmetric. Let, $(c, ..., c)$ denote this equilibrium. From (8), $c \in (a, b)$. 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 (5) 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 $\lambda \geqslant 1$. Consider a sequence $(\lambda^t)_{t=0}^{\infty}$ where $1 < \lambda^t < \lambda^{t-1}$ for all $t \geqslant 1$ 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 (8) $y^t > a$ which, from (4), 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 (5), we have $T(z^t, (n-1)\, y^t + z^t) < 0$. This, together with the fact that $z^t > y^t$ (from (4) and Proposition 1), implies that if $z^t \leqslant c$, 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 (6) and (7). 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. ∎

This result shows that reciprocal preferences can invade a population of materialists under (nonassortative) random matching provided that λ exceeds 1 but is not too high. What is required for this result is that the invader should act spitefully but not too spitefully. The condition that $\lambda > 1$ ensures that a reciprocator in a group of materialists acts spitefully. This spite takes the form of an action higher than the equilibrium for a group of materialists. Negative spillovers lower the materialists' payoffs and strategic substitutability results in their actions falling below the materialist equilibrium level, which in turn raises the reciprocator's payoff above the level that materialists obtain in monomorphic groups. This last claim is only true if the reciprocator's action is not too high, hence the requirement that λ not be too high. 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 find themselves in monomorphic groups when p is close to 1, the payoff to materialists in the population as a whole would exceed that to reciprocators if the latter were too spiteful toward materialists.

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 [32] 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 [32], 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 $\theta = mr$ is a type whose preferences are materialist, and who receives a reciprocator signal. 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 $\rho \in (0, 1)$.[10] 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

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

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 $\Gamma(2)$, 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.[11]

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 $\Gamma \in \mathscr{A}$. There exists $\tilde{\lambda} < n - 1$ such that if $\lambda > \tilde{\lambda}$, a monomorphic population of reciprocators is stable.*

[10] 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 $p\rho_m/(p\rho_m + (1-p)(1-\rho_r))$.

[11] If the global population is 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.

Claim 1. If $\Gamma \in \mathscr{A}$ is separable and satisfies (6), and (7), 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

$$T(d, nd) + \beta_r(n-1)\, H_2(d, nd) \geqslant 0,$$
$$T(d', nd') + \beta_r(n-1)\, H_2(d', nd') \leqslant 0.$$

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 $H_{12} = 0$, (7) implies that $H_{22} < 0$ and hence $H_2(d, nd') > H_2(d, nd)$, a contradiction. ∎

Claim 2. Suppose $\Gamma \in \mathscr{A}$ satisfies (4), (6), (7) and $H_{12} \geqslant 0$, and that $\lambda = n-1$. Then $\Gamma(1)$ and $\Gamma(n)$ have the same (unique) equilibrium.

Proof of Claim 2. For a proof that $\Gamma(n)$ has a unique equilibrium under (6) and (7), see Corchón [9, 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 $T(y, n\bar{x}) \leqslant 0 \leqslant T(z, n\bar{x})$ which from (6) implies that $y \geqslant z$, a contradiction. Similarly, if $y > z$, then $T(z, n\bar{x}) \leqslant 0 \leqslant T(y, n\bar{x})$ which from (6) implies that $z \geqslant y$, 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 (6).

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 $(\lambda^t)_{t=0}^{\infty}$ where $\lambda^{t-1} < \lambda^t < n-1$ for all $t \geqslant 1$ and $\lim_{t \to \infty} \pi^t = n-1$. 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 (4), $y < z$, so $T(y, n\bar{x}) < 0$ is a necessary equilibrium condition. We claim that $c < \bar{x}$. To see this, suppose $\bar{x} \leqslant c$. Then $y < c$ (otherwise we would have $c \leqslant y < z$ contradicting $\bar{x} \leqslant c$), which implies that $T(y, n\bar{x}) > 0 = T(c, nc)$ from (6) and (7), contradicting $T(y, n\bar{x}) \leqslant 0$. Hence $\bar{x} > c$. We next claim that $y < c$. To see this, suppose $c \leqslant y$. Then, since $c < \bar{x}$, (6) and (7) 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 $c \leqslant y$. We have therefore proved that $y < c < \bar{x}$. This, together with (4) and (5), 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, 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.

Propositions 3 and 4, taken together, imply the following. When λ exceeds 1 but is not too large, monomorphic populations of either kind (materialists or reciprocators) are unstable. Materialists can spread in a population consisting largely of reciprocators and vice versa; only interior population states in which both materialists and reciprocators are present can be stable. On the other hand, when λ is sufficiently large, monomorphic

populations of either kind (materialists or reciprocators), are locally stable, and the long-run outcome therefore depends on the initial population composition. In addition, there may exist an intermediate range of values of λ for which exactly one of the two monomorphic population states is stable.

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

5. EFFICIENCY AND ASSORTATIVE INTERACTION

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 [36]). Pure altruism, however, suffers from evolutionary disadvantages under 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 random matching. The following result is a formal statement of the fact that groups of reciprocators outperform groups of materialists.

PROPOSITION 5. *Suppose $\Gamma \in \mathscr{A}$ is separable. 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 (6), (7) and $H_{12} = 0$ together imply that

$$W''(z) = H_{11}(z, nz) + 2nH_{12}(z, nz) + n^2 H_{22}(z, nz) < 0. \tag{10}$$

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 (8), $c \in (a, b)$. 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 $\partial u_i / \partial x_i \geqslant 0$ for all $i \in I$. But at $x = (b, ..., b)$, (8) implies that $T(b, nb) < 0$ and from (5) 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

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

But since $\beta_r \in (0, 1)$ and $H_2 < 0$ from (5), this implies that

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

The above, together with (10), 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

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

for all $\beta \in [0, 1]$ from (6), (7) 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 (5). 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 (10) and the fact that $e = \text{argmax}_{z \in [a, b]} W(z)$, we have $W(d) > W(c)$ as required. (Note that the separability condition $H_{12} = 0$ is

sufficient but not necessary for the above argument; any payoff function H for which inequalities (10) and (11) hold would suffice.) ∎

The above result implies that perfectly assortative interaction favors the growth of reciprocators over groups of materialists. Furthermore, if individuals form groups by voluntary association, the result would be perfectly assortative matching 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 payoffs of materialists, this implies a lower value of their objective function. Consequently, reciprocaters 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.

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 $\bar{\pi}(0) = 0.578 > \bar{\pi}(20) = 0.184 > \bar{\pi}(19) = 0.161$. Hence $\bar{\pi}(k)$ does not decline monotonically with k.

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 (arising from their altruism when they are sufficiently widespread) cause 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 [8], and the references cited therein.)

6. CONCLUSIONS

The analysis in this paper suggests that a population of self-regarding materialists may be unstable in the presence of reciprocal preferences under both nonassortative 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 [16] 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.* [27] for instance.) An

evolutionary analysis that allows for such signalling effects could potentially yield significant new insights.

REFERENCES

1. J. Andreoni and J. H. Miller, "Giving According to GARP: An Experimental Test of the Rationality of Altruism," SSRI Working Paper 9902, University of Wisconsin, 1999.
2. R. Axelrod and W. D. Hamilton, The evolution of cooperation, *Science* **211** (1981), 1390–1396.
3. H. Bester and W. Güth, Is altruism evolutionarily stable, *J. Econ. Behav. Organ.* **34** (1998), 193–209.
4. G. E. Bolton, A comparative model of bargaining: Theory and evidence, *Amer. Econ. Rev.* **81** (1991), 1096–1136.
5. G. E. Bolton, J. Brandts, and A. Ockenfels, Measuring motivations for the reciprocal responses observed in a simple dilemma game, *Exp. Econ.* **1** (1998), 207–219.
6. G. E. Bolton and A. Ockenfels, ERC: A theory of equity, reciprocity and competition, *Amer. Econ. Rev.* **90** (2000), 166–193.
7. S. Bowles and H. Gintis, "The Evolution of Strong Reciprocity," Santa Fe Institute Working Paper, 98-08-073E 1998.
8. J. Canals and F. Vega-Redondo, Multi-level evolution in population games, *Int. J. Game Theory* **27** (1998), 21–35.
9. L. Corchón, "Theories of Imperfectly Competitive Markets," Springer-Verlag, Berlin, 1996.
10. E. Dekel, J. Ely, and O. Yilankaya, Evolution of preferences, mimeo, Northwestern University, 1998.
11. P. Dubey, A. Mas-Colell, and M. Shubik, Efficiency properties of strategic market games, *J. Econ. Theory* **22** (1980), 339–362.
12. M. Dufwenberg and G. Kirchsteiger, "A Theory of Sequential Reciprocity," CentER Discussion paper 9837, Tilburg University, 1998.
13. J. Ely and O. Yilankaya, Evolution of preferences and Nash equilibrium, mimeo, Northwestern University, 1997.
14. A. Falk, E. Fehr, and U. Fischbacher, On the nature of fair behavior, mimeo, University of Zurich, 1997.
15. A. Falk and U. Fischbacher, A theory of reciprocity, mimeo, University of Zurich, 1998.
16. E. Fehr and S. Gächter, Reciprocity and economics: The economic implications of Homo Reciprocans, *Europ. Econ. Rev.* **42** (1998), 845–859.
17. E. Fehr, G. Kirchsteiger, and A. Reidl, Does fairness prevent market clearing? An experimental investigation, *Quart. J. Econ.* **108** (1993), 437–460.
18. E. Fehr and K. M. Schmidt, A theory of fairness, competition, and cooperation, *Quart. J. Econ.* **114** (1999), 817–868.
19. D. Fudenberg and E. S. Maskin, The folk theorem in repeated games with discounting or with incomplete information, *Econometrica* **54** (1986), 533–554.
20. G. Geanakoplos, D. Pearce, and E. Stacchetti, Psychological games and sequential rationality, *Games Econ. Behav.* **1** (1989), 60–79.
21. W. Güth, An evolutionary approach to explaining cooperative behavior by reciprocal incentives, *Int. J. Game Theory* **24** (1995), 323–344.
22. W. Güth, R. Schmittberger, and B. Schwarze, An experimental analysis of ultimatum bargaining, *J. Econ. Behav. Organ.* **3** (1982), 367–388.

23. W. Güth and M. Taari, Explaining reciprocal behavior in simple strategic games: An evolutionary approach, *in* "Explaining Forces and Change: Approaches to Evolutionary Economics" (U. Witt, Ed.), pp. 23–34, University of Michigan Press, Ann Arbor, 1992.
24. S. Huck and J. Oechssler, The indirect evolutionary approach to explaining fair allocations, *Games Econ. Behav.* **28** (1999), 13–24.
25. R. M. Isaac and J. M. Walker, Group size effects in public goods provision: The voluntary contribution mechanism, *Quart. J. Econ.* **103** (1988), 179–200.
26. J. H. Kager, C. Kim, and D. Moser, Fairness in ultimatum games with asymmetric information and asymmetric payoffs, *Games Econ. Behav.* **13** (1996), 100–110.
27. D. Kreps, P. Milgrom, J. Roberts, and R. Wilson, Rational cooperation in the finitely repeated prisoners' dilemma, *J. Econ. Theory* **27** (1982), 245–252.
28. D. Kahneman, J. L. Knetsch, and R. Thaler, Fairness and the assumptions of economics, *J. Bus.* **59** (1986), S285–S300.
29. L. Koçkesen, E. A. Ok, and R. Sethi, The strategic advantage of negatively interdependence preferences, *J. Econ. Theory* **92** (2000), 274–299.
30. L. Koçkesen, E. A. Ok, and R. Sethi, Evolution of interdependent preferences in aggregative games, *Games Econ. Behav.* **31** (2000), 303–310.
31. D. K. Levine, Modeling altruism and spitefulness in experiments, *Rev. Econ. Dynam.* **1** (1998), 593–622.
32. E. A. Ok, and R. Vega-Redondo, On the evolution of individualistic preferences: An incomplete information scenario, *J. Econ. Theory*, forthcoming.
33. M. Rabin, Incorporating fairness into games theory and economics, *Amer. Econ. Rev.* **83** (1993), 1281–1302.
34. A. E. Roth, V. Prasniker, M. Okuno-Fujiwara, S. Zamir, Bargaining and market behavior in Jerusalem, Liubljana, Pittsburgh, and Tokyo: An experimental study, *Amer. Econ. Rev.* **81** (1991), 1068–1095.
35. V. Smith, Microeconomic systems as an experimental science, *Amer. Econ. Rev.* **72** (1982), 923–955.
36. E. Sober and D. S. Wilson, "Unto Others: The Evolution and Psychology of Unselfish Behavior," Harvard University Press, Cambridge, MA, 1998.
37. R. L. Trivers, The evolution of reciprocal altruism, *Quart. Rev. Biol.* **46** (1971), 35–57.