J. theor. Biol. (2001) **209,** 87–95
doi:10.1006/jtbi.2000.2248, available online at http://www.idealibrary.com on IDEAL®

Heterogeneity Stabilizes Reciprocal Altruism Interactions

MICHAEL A. FISHMAN*, ARNON LOTEM AND LEWI STONE

Department of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, *Israel*

(*Received on* 25 *October* 1999, *Accepted in revised form on* 9 *December* 2000)

In considering the phenomena of reciprocal altruism few would dispute that there are differences in individual quality—in particular, that for some individuals, at least on occasion, the cost of doing favors will exceed the potential of future benefits. That is, at any given time, a typical population is heterogeneous with respect to the *affordability* of reciprocal altruism. However, methodological limitations of the traditional analytical framework—*Single Type* (*symmetric*) *Evolutionary Game Theory*—have restricted previous analytical efforts to addressing populations idealized in terms of their averages. Here we use the methods of *Multitype Evolutionary Game Theory* to analyse the role of individual differences in direct reciprocity interactions. Multitype analysis shows that non-idealized populations possess an ESS profile wherein individuals who cannot afford reciprocity (low-quality) defect, while individuals who derive net benefits from reciprocity (high-quality) cooperate. Furthermore, this cooperation is implemented via unmodified tit-for-tat (TfT) strategy. Hence, our results may help resolve a long-standing problem concerning the evolutionary stability of TfT in direct reciprocal altruism. Finally, this difference between idealized and real populations is not restricted to direct reciprocal cooperation. Previously (Lotem *et al.*, 1999) we have demonstrated evolutionarily stable *indirect reciprocal cooperation* among high-quality individuals in heterogeneous populations.

© 2001 Academic Press

0. Introduction

Altruistic behavior is commonly attributed to inclusive fitness, or reciprocity (Krebs & Davies, 1993). The idea of *reciprocal altruism*, albeit one motivated by conscious calculation, was first proposed by Darwin: "as the reasoning powers and foresight … become improved, each man would soon learn from experience that if he aided his fellow-men, he would commonly receive aid in return". (Darwin, 1871). The idea of reciprocal altruism in its modern form, i.e. as an unconscious inheritable trait maintained by Darwinian fitness advantages, was first advanced by Williams (1966), and given rigor by Trivers (1971).

By using game theoretical analysis Trivers has shown that an individual may help an unrelated conspecific whenever: (i) the cost to the donor is less than the benefit to the recipient, and (ii) the favor is likely to be returned at a latter date. Hence, in order for reciprocal altruism (cooperation) to persist, cooperators must protect themselves from exploitation by individuals who accept favors, but do not reciprocate—*defectors*. Trivers, aided by W. D. Hamilton (Trivers, 1971, p. 39), resolved this issue by postulating that cooperation is conditional, i.e. a cooperator will

* Author to whom correspondence should be addressed. E-mail: mafish@post.tau.ac.il

0022–5193/01/050087 + 09 $35.00/0

© 2001 Academic Press

keep helping an unrelated individual, unless the latter refuses to reciprocate.† These theoretical results soon gained empirical support from observations detailing reciprocal exchange of favors in group-living species (Fisher, 1980; Seyfarth & Cheney, 1984; Wilkinson, 1984).

The concept of helpfulness conditional upon opponents' reciprocity was further developed by Axelrod and co-workers (Axelrod & Hamilton, 1981; Axelrod, 1984; Trivers, 1985) resulting in the formulation of *tit-for-tat* (TfT) strategy: a TfT player will punish defection by refusing help in turn, but otherwise will cooperate. In particular, these authors demonstrated that a population of TfT players cannot be invaded by defectors, i.e. TfT is an *Evolutionarily Stable Strategy* (ESS) in the context of the TfT vs. defector contest. However, Selten & Hammerstein (1984) have shown that TfT is not an ESS under biologically plausible conditions. *Briefly*: mutants that lost the ability to make TfT-type evaluation of opponents, and thus help others unconditionally, can increase through *genetic drift* in a population of TfT players, and a population containing substantial fraction of these *unconditional altruists* (UA) can be invaded by defectors.

The demonstration that TfT is not an ESS led to the concerted effort to formulate cooperation strategies superior to TfT in the sense of not being subject to invasion by unconditional altruists, reviewed by (Nowak *et al*., 1995; Brembs, 1996). The alternate approach of this paper is based on the following considerations. Even if the benefits of receiving help are the same for all members of a population, there are differences (genetic or phenotypic) among individuals with respect to the ability to donate help. Hence, for any type of exchange of favors, there are three possible situations.

(a) The costs of reciprocity are less than its benefits for all individuals involved.
(b) The costs of reciprocity exceed its benefits for all individuals involved.
(c) The costs of reciprocity exceed its benefits for some individuals, and are less than these benefits for the rest. Hence, at any given time, the subject population consists of two *quality classes*: *low-quality* individuals—who cannot "afford" reciprocity vs. *high-quality* individuals—who derive net benefits by exchanging favors.

By adopting a representation in terms of population averages, cases (a) and (b) can be analysed in terms of symmetric (single-type) evolutionary game theory (cf. Nowak *et al*., 1995; Brembs, 1996), which addresses situations where all contestants have the same choice of game strategies and receive the same payoffs for any particular interaction (Maynard Smith, 1982). However, such a simplification would reduce case (c) to either (a) or (b), and therefore is inappropriate. Thus, to analyse case (c), we must use the methods of *Multitype Evolutionary Game Theory* that allow analysis of contests between contestants differing in the strategy choices available to them, or having the same strategy choices but different payoffs for some of the possible interactions (Cressman, 1992; Weibull, 1996).

Our multitype analysis shows that individuals who cannot afford to reciprocate (e.g. young, sick, handicapped, or those that simply do not have sufficient resources at a given time) will defect by default. Thus, these *phenotypic defectors* (Lotem *et al*., 1999) are a special case of *phenodeviation*—a name for the disruptive effects of the environment on genotype expression proposed by Thornhill & Møller (1997) in their seminal work on developmental stability. Because defection by default is a phenodeviation, and thus cannot be eliminated by natural selection, it is a persistent feature of real populations.

The persistence of defection confers an advantage on TfT players *vis-a-vis* unconditional altruists. This advantage is absolute, i.e. TfT players always have higher fitness than unconditional altruists, leading to the elimination of the latter from the population. As was shown previously (Axelrod & Hamilton, 1981), in the absence of

†Trivers used a metaphor, known as the *Prisoner's Dilemma* (PD), from game theory. In this game, the two players have the choice of cooperating or defecting. The payoff to each when they cooperate is greater than the payoff for mutual defection, but less than the payoff to a defector playing against a cooperator. Finally, the payoff to a cooperator playing against a defector is the least of all. A single-stage PD game can be shown to have a unique stable solution—mutual defection (cf. Fudenberg & Tirole, 1996, Section 1.1.3). Trivers have shown that in an open-ended series of PD games between two opponents—a *Repeated Prisoner's Dilemma* (RPD), a conditional cooperator does better than a defector.

unconditional altruists, TfT is an ESS. Thus, we arrive at an ESS profile in which individuals who derive net benefits from reciprocity play TfT, whereas individuals that cannot afford reciprocity defect.

Since an individual's ability to reciprocate (quality) changes with time, our results can be interpreted in two, not necessarily mutually exclusive, ways. (i) A population might be permanently divided into high-quality cooperators and low-quality defectors. (ii) Individuals can switch behavior as their capacity for reciprocity (quality) varies. In this latter interpretation, our ESS result represents cooperators who occasionally defect. Thus, our derived strategy profile is reminiscent of the evolutionarily stable "mistake-making TfT" strategy proposed by Boyd (1989).

Finally, the stabilizing effect of phenotypic defection, is not restricted to direct-reciprocity interactions. Previously (Lotem *et al.*, 1999), we have shown that the presence of phenotypic defectors, introduced as a modelling assumption, stabilizes cooperation in the analogous situations of *indirect reciprocity* interactions (Nowak & Sigmund, 1998a, b).

This paper is organized as follows. In Section 1, we introduce a notation, one that we find convenient for multitype analysis, and recapitulate the work of our predecessors in this notation. That is, we formulate a symmetric game theoretical model addressing cases (a) and (b), and show that in these situations defector strategy is the unique ESS. In Section 2, we extend the single-type model of Section 1 to a multitype model of the two quality classes of situation (c). To facilitate presentation we confine some of the technical details of the analysis to the appendix.

1. Symmetric Model (Idealized Populations)

We start the analysis by constructing a symmetric game theoretical model for idealized populations. That is, we address the issue of direct reciprocal altruism in terms of population averages. This yields situations (a) or (b), i.e. the (average) costs of reciprocity are either less (a), or greater than (b) its benefits. As discussed in the introduction, we consider three evolutionary game strategies (heritable behavior phenotypes): unconditional altruists (UA) that help others indiscriminately; defectors (DE) that solicit, but never donate help; and conditional altruists, or TfT players—who retaliate for each defection by refusing help in the next interaction, but otherwise act as unconditional altruists. Thus, unlike the invariant responses of the UA and DE players, the response to a request for help by a TfT player depends on its memory of previous interactions. That is, a TfT player always helps unconditional altruists and other TfT players, but helps defectors only when it lacks information to classify them. Denoting the probability that an individual requesting help has been requested to help recently enough for its response to be remembered by $0 < r < 1$, we obtain the TfT response scheme summarized in Fig. 1.

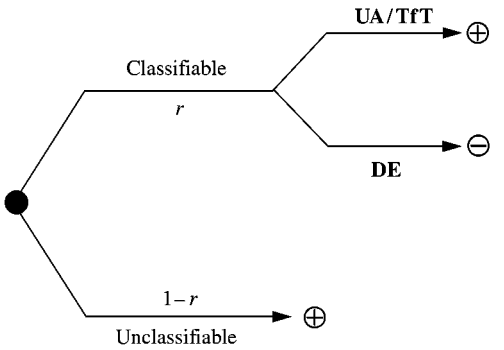

FIG. 1. Here $0 < r < 1$ is the probability that an individual requesting help has been requested to help recently enough for its response to be remembered: defection is punished with a refusal to help (−), and cooperation is rewarded with cooperation (+). Alternatively, with probability $1 - r$, there is no definite memory of previous interaction, and therefore the **TfT** is not motivated to refuse help. Thus, the value of r depends on the probability of repeated interactions and the fidelity of memory/individual recognition. Considered in a different way, r is the probability that the favor will be repaid, if the recipient is another **TfT**.

Let us denote the average (*per capita*) accumulated benefits of receiving help over a lifetime by B, and the average lifelong costs of donating help by C; we use capital letters to distinguish these, per lifespan, payoffs from the per encounter payoffs more usual in the literature (cf. Nowak & Sigmund, 1998b).‡

‡ If the per encounter benefits and costs are given by b and c, respectively; the probability of t encounters per lifespan is given by w^t $(0 < w < 1)$; and, on the average, individuals take turns soliciting and being solicited for help: then $B = b/(2(1 - w))$ and $C = c/(2(1 - w))$.

In these terms the payoff matrix for donating help is given by

$$\begin{array}{c} \begin{matrix} \mathbf{UA} & \mathbf{TfT} & \mathbf{DE} \end{matrix} \\ \begin{pmatrix} -C & -C & -C \\ -C & -C & -(1-r)C \\ 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA} \\ \mathbf{TfT} \\ \mathbf{DE} \end{matrix}. \quad (1a)$$

Note that the entry *i–j* represents the costs of help given by (an average player of) strategy-*i* to (an average player of) strategy-*j*. Thus, to calculate the benefits of receiving help, we transpose matrix (1a) and substitute $+B$ for $-C$. This yields

$$\begin{array}{c} \begin{matrix} \mathbf{UA} & \mathbf{TfT} & \mathbf{DE} \end{matrix} \\ \begin{pmatrix} B & B & 0 \\ B & B & 0 \\ B & (1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA} \\ \mathbf{TfT} \\ \mathbf{DE} \end{matrix}. \quad (1b)$$

The payoff matrix, P, for both giving and receiving help, the *game matrix*, is obtained by adding matrices (1a, b) to obtain

$$P = \begin{array}{c} \begin{matrix} \mathbf{UA} & \mathbf{TfT} & \mathbf{DE} \end{matrix} \\ \begin{pmatrix} B-C & B-C & -C \\ B-C & B-C & -(1-r)C \\ B & (1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA} \\ \mathbf{TfT} \\ \mathbf{DE} \end{matrix}. \quad (2)$$

As discussed above, in this section we consider two possibilities.

(a) *The costs of reciprocity are less than its benefits, i.e.* $C < B$. In the appendix we show that, if $C < B$, then system (2) has a unique ESS solution, **DE**, i.e. defectors displace individuals using alternative strategies, resulting in a population consisting of defectors only.

(b) *The costs of reciprocity are greater than its benefits, i.e.* $C > B$. If $C > B$, then every element of the third row of P is greater than the corresponding elements of its first and second rows. That is, at any composition of the population, defectors have higher fitness than cooperators (**UA** or **TfT**), leading to the elimination of the latter from the population. Formally, **UA** and **TfT** are *strictly dominated* by **DE**, and can be excluded, i.e. a reduced system—obtained by excluding the strictly dominated strategies—has the same ESS solutions as the original system (cf. Weibull, 1996, Chapter 3.2.1). Therefore, defection is again the unique ESS.

2. A Multitype Model for Heterogeneous Populations

In this section, we analyse a situation where the costs of reciprocity exceed its benefits for some individuals, and are less than these benefits for the rest. We start by dividing the population into two classes: *low-quality* individuals for whom costs of reciprocity exceed its benefits vs. *high-quality* individuals for whom reciprocity yields net benefits. The membership in a class is not necessarily hereditary—a reader might find it convenient to think of these quality classes as juveniles and mature individuals, respectively. We shall denote the frequency of low-quality individuals by $0 < q < 1$ (cases $q = 0, 1$ have been addressed in the previous section). We retain C as the *average* for the accumulated lifelong costs of altruism in the high-quality class, and denote the corresponding value for the low-quality class by D. We retain the use of B for the accumulated lifelong benefits. As discussed above, $C < B < D$. Using r as in Section 1, and using the subscripts H and L to denote the quality classes, we have the following payoff matrices: $P_{HH}, P_{HL}, P_{LH}, P_{LL}$. Here the first subscript defines the focal (recipient of the payoff, the row strategy) and the second subscript defines the opponent:

$$P_{HH} = (1-q) \begin{array}{c} \begin{matrix} \mathbf{UA}_H & \mathbf{TfT}_H & \mathbf{DE}_H \end{matrix} \\ \begin{pmatrix} B-C & B-C & -C \\ B-C & B-C & -(1-r)C \\ B & (1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA}_H \\ \mathbf{TfT}_H \\ \mathbf{DE}_H \end{matrix}, \quad (3)$$

$$P_{HL} = q \begin{array}{c} \begin{matrix} \mathbf{UA}_L & \mathbf{TfT}_L & \mathbf{DE}_L \end{matrix} \\ \begin{pmatrix} B-C & B-C & -C \\ B-C & B-C & -(1-r)C \\ B & (1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA}_H \\ \mathbf{TfT}_H \\ \mathbf{DE}_H \end{matrix}, \quad (4)$$

$$P_{LH} = (1-q)\begin{array}{c} \begin{matrix} \mathbf{UA}_H & \mathbf{TfT}_H & \mathbf{DE}_H \end{matrix} \\ \begin{pmatrix} B-D & B-D & -D \\ B-D & B-D & -(1-r)D \\ B & (1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA}_L \\ \mathbf{TfT}_L, \\ \mathbf{DE}_L \end{matrix} \tag{5}$$

$$P_{LL} = q\begin{array}{c} \begin{matrix} \mathbf{UA}_L & \mathbf{TfT}_L & \mathbf{DE}_L \end{matrix} \\ \begin{pmatrix} B-D & B-D & -D \\ B-D & B-D & -(1-r)D \\ B & (1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA}_L \\ \mathbf{TfT}_L. \\ \mathbf{DE}_L \end{matrix} \tag{6}$$

Note that the payoffs depend on the frequencies of the two quality types in the population. For example, every element of the P_{HH} and P_{LH} is multiplied by $(1-q)$ because this is the probability to encounter a high-quality opponent.

The general mathematical framework for analysing evolutionary stability in games with two types of players was developed by Cressman and co-workers (Cressman & Dash, 1991; Cressman, 1992). The specific case of system (3–6), however, can be analysed by taking advantage of the fact that, similar to the case for symmetric games discussed above, strictly dominated strategies in multitype games can be excluded without affecting the ESS solutions of the game (cf. Weibull, 1996, Section 5.6.1). To wit, since $B < D$, every element of the third row of P_{LH} and P_{LL} is greater than the corresponding elements of the first and second rows. Hence, $\mathbf{UA}_L$ and $\mathbf{TfT}_L$ are *strictly dominated* by $\mathbf{DE}_L$ and can be excluded. That is, we see that low-quality individuals "must" defect. Exclusion of $\mathbf{UA}_L$ and $\mathbf{TfT}_L$ yields a reduced system:

$$(1-q)\begin{array}{c} \begin{matrix} \mathbf{UA}_H & \mathbf{TfT}_H & \mathbf{DE}_H \end{matrix} \\ \begin{pmatrix} B-C & B-C & -C \\ B-C & B-C & -(1-r)C \\ B & (1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA}_H \\ \mathbf{TfT}_H, \\ \mathbf{DE}_H \end{matrix}$$

$$-qC\begin{array}{c} \mathbf{DE}_L \\ \begin{pmatrix} 1 \\ 1-r \\ 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA}_H \\ \mathbf{TfT}_H, \\ \mathbf{DE}_H \end{matrix} \tag{7}$$

$$\begin{array}{cccc} \mathbf{UA}_H & \mathbf{TfT}_H & \mathbf{DE}_L & \mathbf{DE}_L \\ ((1-q)B & (1-r)(1-q)B & 0) & \mathbf{DE}_L, \quad (0) \\ & & & \mathbf{DE}_L. \end{array}$$

Since all low-quality individuals are (phenotypic) defectors, the ESS solutions of system (7), and hence system (3–6), have the form

$$(\mathbf{x}^*, \mathbf{DE}_L), \tag{8}$$

where $\mathbf{x}^*$ is an ESS solution of the system obtained by combining the payoffs for interacting with $\mathbf{DE}_L$ to the payoffs for high-quality *vs.* high-quality interactions. That is, we add the row elements of the reduced P_{HL} to each element of the appropriate row of P_{HH} to obtain

$$\begin{array}{c} \begin{matrix} \mathbf{UA}_H & \mathbf{TfT}_H & \mathbf{DE}_H \end{matrix} \\ \begin{pmatrix} (1-q)B-C & (1-q)B-C & -C \\ (1-q)B-(1-qr)C & (1-q)B-(1-qr)C & -(1-r)C \\ (1-q)B & (1-q)(1-r)B & 0 \end{pmatrix} \end{array} \begin{matrix} \mathbf{UA}_H \\ \mathbf{TfT}_H. \\ \mathbf{DE}_H \end{matrix} \tag{9}$$

Since $q, r > 0$, every element of the second row is greater than the corresponding element of the first row. This is due to the fact that the burden imposed by the presence of phenotypic defectors on unconditional altruists is greater than the corresponding burden, eqn (7), on the TfT players. Formally, TfT_H strictly dominates $\mathbf{UA}_H$—and therefore, as discussed above, we can exclude $\mathbf{UA}_H$. This yields a reduced system

$$P_H = \begin{pmatrix} p_{tt} & p_{td} \\ p_{dt} & p_{dd} \end{pmatrix}$$

$$= \begin{array}{c} \\ \\ \end{array} \begin{array}{c} \mathbf{TfT}_H \\ \\ \end{array} \begin{array}{c} \mathbf{DE}_H \\ \\ \end{array} \begin{pmatrix} (1-q)B-(1-qr)C & -(1-r)C \\ (1-q)(1-r)B & 0 \end{pmatrix} \begin{array}{l} \mathbf{TfT}_H \\ \mathbf{DE}_H \end{array}. \tag{10}$$

We see that $p_{dd} > p_{td}$, i.e. defector playing against defector does better than a TfT playing against defector. Hence, a population of defectors cannot be invaded by TfT players. Consequently, defection is an evolutionarily stable strategy of system (10).

If $p_{dt} > p_{tt}$ as well, then defector playing against TfT does better than a TfT playing against TfT (exploitation pays better than cooperation). Hence, a population of TfT players can be invaded and taken over by defectors. Thus, if $p_{dt} > p_{tt}$, defection is the only ESS of system (10). However, if $p_{tt} > p_{dt}$, then cooperation pays better than exploitation. In particular, a population of TfT players cannot be invaded by defectors. That is, as discussed in connection with eqn (8), system (3–6) has two ESS solutions ($\mathbf{DE}_H$, $\mathbf{DE}_L$) or ($\mathbf{TfT}_H$, $\mathbf{DE}_L$).

Now,

$$\begin{aligned} p_{tt} - p_{dt} &= (rB - C) - qr(B - C) \\ &= rB(1-\rho)(\theta - q), \end{aligned} \tag{11}$$

where

$$\rho = \frac{C}{B} \quad \text{and} \quad \theta = \frac{rB - C}{r(B - C)} = \frac{1}{r}\frac{r-\rho}{1-\rho}.$$

Since $C < B$, $\rho < 1$. Thus, for cooperation to be more productive than exploitation, we must have $q < \theta$. Since $q > 0$, we must have $\theta > 0$, which in turn requires $r > \rho$, i.e. $rB > C$. As discussed in Section 1, rB is the expectation of repayment when dealing with another TfT player. Hence, we obtain the unsurprising conclusion that *cooperation can persist only if the expected repayment exceeds the investment*.§

§This result is analogous to the *Hamilton's rule of kin altruism* (Hamilton, 1964). According to Hamilton's rule help may be donated to a relative if the degree of relatedness times the benefit to the recipient (the inclusive fitness benefits to the donor) exceeds the donor's costs. That is, in both cases apparently altruistic acts are undertaken only when they yield net benefits to the "altruist" (Nowak & Sigmund, 1998b).

Condition $rB > C$ is both necessary and sufficient when we consider competition between TfT players and defectors (Axelrod & Hamilton, 1981). However, as detailed in Section 1 and in the appendix: eqns (A5, A6), because in the absence of defectors there is no difference in fitness between TfT and UA players—TfT playing populations can be invaded by unconditional altruists and a subsequent mixed population can be invaded by defectors. This result, however, is only obtained when we neglect the heterogeneity of real populations. In heterogeneous populations, there are individuals who cannot afford reciprocity, eqns (3–6), and therefore defect by default—*phenotypic defectors*. In the presence of these phenotypic defectors, unconditional altruists have lower fitness than TfT players, eqn (9), and can be excluded. Thus, in heterogeneous populations, the situation reduces to the competition between (high-quality) TfT players and (high-quality) defectors, modified by the presence of phenotypic defectors.

Although phenotypic defectors prevent destabilization of cooperation by unconditional altruists, their presence is not an unmixed blessing. Because TfT players help defectors (in particular, phenotypic defectors) when in doubt (Fig. 1), the presence of phenotypic defectors imposes a burden on TfT players, and decreases their ability to compete with defectors. Thus cooperation persists only if the frequency of phenotypic defectors (q) is less than θ, i.e. θ can be thought of as the *tolerance capacity* (in analogy with the *carrying capacity* term of the logistic equation), for the burden of phenotypic defectors. We summarize these results in Fig. 2.

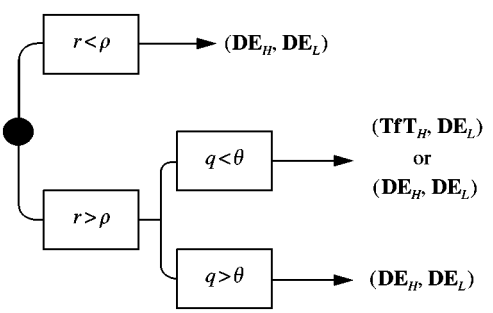

FIG. 2. Here the inequalities on the paths from the origin (●) to endpoints represent the conditions that must be satisfied if the strategy profile(s) at the endpoint to be ESS.

Thus, we see that cooperation is possible if a population contains individuals who cannot afford reciprocity, but the frequency of such, low-quality individuals is not too high.

3. Discussion

The analysis of reciprocity undertaken in this paper shows that reciprocal altruism can be stable when individual variation is taken into account. Genetic and/or phenotypic differences in ability among individuals, in particular in individual costs for donating help, create three distinct situations. When the costs of reciprocity are either less than its benefits or exceed its benefits for all individuals involved, then the situation can be analysed in terms of the single-type (symmetric) evolutionary game theory, and reciprocity is not an ESS (Selten & Hammerstein, 1984). However, the most likely situation is when the costs of reciprocity exceed the benefits for some individuals, but are less than the benefits for the rest. That is, the subject population consists of two *quality classes*: *low-quality* individuals—who cannot "afford" reciprocity vs. *high-quality* individuals—who derive net benefits from reciprocity, and must be analysed using the methods of the multitype evolutionary game theory (Cressman, 1992). Multitype analysis shows that, given the appropriate conditions, such a *two-class* population exhibits an ESS profile wherein individuals who cannot afford reciprocity defect, while individuals who derive net benefits from reciprocity cooperate by playing unmodified TfT.

The stabilizing effect of individual variation derives from the fact that individuals who cannot afford reciprocity defect by default. The persistent presence of these *phenotypic defectors* abrogates the ability of the unconditional altruist mutants to invade TfT populations—see Lotem *et al.* (1999) for similar effects in indirect reciprocity. Thus, contrary to the results in idealized populations, persistent cooperation based on unmodified TfT strategy is possible in real (heterogeneous) populations.

As discussed in the introduction, a large number of conditional cooperation strategies able to discriminate against unconditional altruists (cf. Nowak *et al.*, 1995; Brembs, 1996), were formulated following the demonstration that TfT playing populations can be invaded by unconditional altruists and subsequently by defectors (Selten & Hammerstein, 1984). Our work demonstrates that TfT is evolutionarily stable in competition with unconditional strategies in heterogeneous populations. However, it is by no means certain that an analogous analysis of the competition between TfT and some of the more sophisticated conditional cooperation strategies will demonstrate domination by the former. Thus, examination of the effects of heterogeneity on the functioning of conditional cooperation strategies, and of their relative merits in heterogeneous context, though beyond the scope of the present paper, is the next logical step in game theoretical investigation of reciprocity.

Above all else, the current study illustrates that individual variation is more than just a noise, and thus the study of the evolution of behavior in terms of population averages may yield misleading results. In terms of mathematical methods, our results highlight the usefulness of multitype evolutionary game theory in analysing real (heterogeneous) populations. On the empirical level, they illustrate the importance of studying variations in quality in relation to the behavioral phenotypes.

We would like to use this opportunity to thank the unknown referees for encouraging evaluation and constructive criticism.

REFERENCES

AXELROD, R. (1984). *The Evolution of Cooperation* (reprinted 1989). Harmondsworth: Penguin.

AXELROD, R. & HAMILTON, W. D. (1981). The evolution of cooperation. *Science* **242,** 1390–1396.

BOYD, R. (1989). Mistakes allow evolutionary stability in the Repeated Prisoner's Dilemma game. *J. theor. Biol.* **136,** 47–56.

BREMBS, B. (1996). Chaos, cheating and cooperation: potential solutions to the Prisoner's Dilemma. *OIKOS* **76,** 14–24.

CRESSMAN, R. (1992). *The Stability Concept of Evolutionary Game Theory*. Berlin: Springer-Verlag.

CRESSMAN, R. & DASH, A. T. (1991). Strong stability and evolutionarily stable strategies with two types of players. *J. Math. Biol.* **30,** 89–99.

DARWIN, C. (1871). *The Decent of Man, and Selection in Relation to Sex* (reprinted 1981). Princeton, NJ: Princeton University Press.

FISHER, E. A. (1980). The relationship between mating systems and simultaneous Hermaphroditism in the coral reef fish *Hypoplectrus nigricans*. *Anim. Behav.* **28,** 620–633.
FUDENBERG, D. & TIROLE, J. (1996). *Game Theory*, *5th Printing*. Cambridge, MA: MIT Press.
HAMILTON, W. D. (1964). The genetic evolution of social behavior. I & II. *J. theor. Biol.* **7,** 1–52.
KREBS, J. R. & DAVIES, N. B. (1993). *An Introduction to Behavioral Ecology*: §11, 3rd Edn. Oxford: Blackwell Scientific Publications.
LOTEM, A., FISHMAN, M. A. & STONE, L. (1999). Evolution of cooperation between individuals. *Nature* **400,** 226–227.
MAYNARD SMITH, J. (1982). *Evolution and the theory of Games*. Cambridge: Cambridge University Press.
NOWAK, M., MAY, R. M. & SIGMUND, K. (1995). The arithmetics of mutual help. *Sci. Am.* **272,** 76–81.
NOWAK, M. & SIGMUND, K. (1998a). Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573–577.
NOWAK, M. & SIGMUND, K. (1998b). The dynamics of indirect reciprocity. *J. theor. Biol.* **194,** 561–574.
SELTEN, R. & HAMMERSTEIN, P. (1984). Gaps in Harley's argument on evolutionary stable learning rules and in the logic of TfT. *Behav. Brain. Sci.* **7,** 115–116.
SEYFARTH, R. M. & CHENEY, D. I. (1984). Grooming, alliances and reciprocal altruism in vervet monkeys. *Nature* **308,** 541–543.
THORNHILL, R. & MØLLER, A. P. (1997). Developmental stability, disease and medicine. *Biol. Rev.* **72,** 497–548.
TRIVERS, R. (1971). The evolution of reciprocal altruism. *Quart. Rev. Biol.* **46,** 35–56.
TRIVERS, R. (1985). *Social Evolution*. Menlo Park, CA: Benjamin/Cummings.
WEIBULL, J. W. (1996). *Evolutionary Game Theory*, 2nd Edn. Cambridge, MA: MIT Press.
WILKINSON, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature* **308,** 181–184.
WILLIAMS, G. C. (1966). *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton, NJ: Princeton University Press.

Appendix

Methods and Notation

Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ be the standard basis of $\Re^3$, and let us denote the strategy set for the game given by eqn (2) (payoff matrix P) by

$$\mathbf{X} = \{x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3 | x_1, x_1, x_3 \geqslant 0 \quad \text{and} \quad x_1 + x_1 + x_3 = 1\}. \tag{A1}$$

That is, $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ represent the *pure strategies* **UA**, **TfT**, and **DE**, and their convex combinations represent *mixed strategies*. In these terms, the payoff for playing strategy $\mathbf{x} \in \mathbf{X}$ against a strategy $\mathbf{y} \in \mathbf{X}$ is given by

$$u(\mathbf{x}, \mathbf{y}) \equiv \mathbf{x}^{\mathrm{t}} \circ \mathrm{P} \circ \mathbf{y} = (1 - y_3 - rx_3y_2)B - (1 - x_3 - rx_2y_3)C. \tag{A2}$$

The ESS criterion (Cressman, 1992, Section 2) is given by

Strategy $\mathbf{x}^* \in \mathbf{X}$ is an ESS if for any $\mathbf{x} \in \mathbf{X}$, $\mathbf{x} \neq \mathbf{x}^*$ implies

$$\Gamma_0(\mathbf{x}^*, \mathbf{x}) \equiv u(\mathbf{x}^* - \mathbf{x}, \mathbf{x}^*) > 0 \quad \text{or}$$

$$\Gamma_0(\mathbf{x}^*, \mathbf{x}) = 0$$

and

$$\Gamma_1(\mathbf{x}^*, \mathbf{x}) \equiv u(\mathbf{x}^* - \mathbf{x}, \mathbf{x} - \mathbf{x}^*) > 0. \tag{A3}$$

Evolutionary games do not necessarily have an ESS point solution. There is also the possibility of a set of solutions exhibiting neutral stability among themselves, while being ESS like in comparison with the strategies not in that set—*evolutionarily stable sets* (ES sets). Formally (Cressman, 1992, Chapter 6), a proper subset $\Lambda \subset \mathbf{X}$ is an ES set if every $\mathbf{x}_\lambda \in \Lambda$ satisfies condition (A3) versus every $\mathbf{x} \notin \Lambda$, while being neutrally stable relative to other strategies in Λ.

Obviously, we cannot check every strategy in the continuous set $\mathbf{X}$ for being an ESS (let alone an element of an ES set). Hence, we enumerate the ESS solutions as a two-step process. First, we find all the potential ESS solutions by using the fact that every ESS solution is a *Nash Equilibrium* point (though not vice versa) and therefore if $\mathbf{x}^* \in \mathbf{X}$ is an ESS, then $u(\mathbf{e}_j, \mathbf{x}^*) = u(\mathbf{x}^*, \mathbf{x}^*)$ or $x_j^* = 0, \forall j$.

Consequently, every solution of the system of equations

$$[u(\mathbf{e}_j - \mathbf{x}, \mathbf{x})]x_j = 0 \quad \forall j \tag{A4}$$

on $\mathbf{X}$ is a potential ESS. Once we have derived all the potential ESS solutions, we apply the ESS criterion (A3).

In the specific case of system (2), there is one potential ES set solution and two potential ESS solutions.

The potential ES solution

$$\Lambda = \{\lambda \mathbf{e}_1 + (1 - \lambda)\mathbf{e}_2 \,|\, \lambda \in [0, 1]\} \qquad \text{(A5)}$$

represents the state where defectors are absent, and therefore there is no difference between unconditional altruists and TfT players. Since

$$\mathbf{e}_1 \in \Lambda \quad \text{and} \quad \mathbf{e}_3 \notin \Lambda$$

but

$$\Gamma_0(\mathbf{e}_1, \mathbf{e}_3) = -C. \qquad \text{(A6)}$$

Λ is not an *evolutionary stable set*.

Next, we have a potential ESS solution wherein TfT players and defectors coexist

$$\mathbf{x}_2 = \xi \mathbf{e}_2 + (1 - \xi)\mathbf{e}_3\text{:} \quad \xi = \frac{(1 - r)C}{r(B - C)}. \qquad \text{(A7)}$$

However, since $\mathbf{x}_2 \neq \mathbf{e}_2$ but

$$\Gamma_0(\mathbf{x}_2, \mathbf{e}_2) = 0 \quad \text{and}$$

$$\Gamma_1(\mathbf{x}_2, \mathbf{e}_2) = -r(B - C)(1 - \xi)^2. \qquad \text{(A8)}$$

$\mathbf{x}_2$ is not an ESS.

Finally, we have a potential ESS solution wherein defectors displace both unconditional altruists and TfT players

$$\mathbf{x}_3 = \mathbf{e}_3 \quad \text{and} \quad \Gamma_0(\mathbf{x}_3, \mathbf{x}) = C[x_1 + (1 - r)x_2]. \qquad \text{(A9, A10)}$$

Hence, since $0 < r < 1$, $\mathbf{x}_3$(**DE**) is an ESS for all parameter values.