Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 231 (2004) 475–486
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

The logic of reprobation: assessment and action rules for indirect reciprocation

Hannelore Brandt[a], Karl Sigmund[a,b,*]

[a]*Fakultät für Mathematik, Nordbergstrasse 15, 1090 Wien, Austria*
[b]*Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, 2361 Laxenburg, Austria*

Received 6 February 2004; received in revised form 4 June 2004; accepted 7 June 2004
Available online 25 September 2004

Abstract

Ever since image-based models for indirect reciprocity were introduced, the relative merits of scoring vs. standing have been discussed to find out how important it is to differentiate between justified and non-justified defections. This is analogous to the question whether punishment can sustain cooperation even when it is costly. We show that an answer to this question can depend on details of the model, for instance concerning the probability distribution of the number of interactions experienced per player. We use extensive individual-based simulations to compare scoring, standing and other forms of assessing defections, and show that several forms of indirect reciprocation can robustly sustain cooperation. By most standards, standing is better than scoring, but nevertheless scoring is able to sustain cooperation in the presence of errors. The model presented here is based on three specifications: each player has a personal list of images of all co-players, a specific way of judging an observed situation, and a specific strategy to decide whether to cooperate or not.

Keywords: Indirect reciprocity; Evolution of cooperation; Image-scoring

1. Introduction

There is a whiff of paradox around the concept of indirect reciprocation. With direct reciprocation, the help which a donor provides for a recipient will eventually be returned by that recipient, so that if the benefit b for the recipient exceeds the cost c for the donor, both parties obtain a net gain (Trivers, 1971). Indirect reciprocation occurs if the help is eventually returned, i.e. the donor is compensated, but by a third party, rather than by the recipient. This system seems even more vulnerable to selfish exploitation than direct reciprocation. Indeed, common sense and theoretical models alike show that if two players interact often enough, the threat of stopping to help whenever the co-player fails to return the favour can ensure mutual help: in other words, retaliatory strategies can foster cooperation in an iterated Prisoner's Dilemma game (Axelrod and Hamilton, 1981). But in the context of indirect reciprocation, the same two players never meet twice. How, then, can defectors be punished? Their victims cannot take them to account. And yet, as many economic experiments show, indirect reciprocation works in human societies (Wedekind and Milinski, 2000; Milinski et al., 2001, 2002; Seinen and Schram, 2001; Bolton et al., 2001; Wedekind and Braithwaite, 2002; Camerer, 2003), and it has even been touted to provide the biological basis for our morality (Alexander, 1987).

Nowak and Sigmund (1998a, b) have suggested a theoretical model based on the concept of a score (an abstract measure for the reputation of being a person who gives help). The score of potential donors increases

*Corresponding author. Faculty for Mathematics, Universitat Wien, Nordbergstrasse 15, Wien 1090, Austria. Tel.: +43-1-4277-50612; fax: +43-1-4277-9506.
E-mail address: karl.sigmund@univie.ac.at (K. Sigmund).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.06.032

if they actually do provide help, and decreases if they refuse to do so. Numerical simulations show that discriminating strategies, providing help if the recipient's score exceeds some given threshold, can indeed evolve and lead to cooperative societies. The score provides an instrument for channeling help towards those who help, and thus to suppress defectors. But by discriminating against low-scorers, players lower their own score, and therefore risk being refused help on a later occasion. Can they be expected to discriminate if this hurts themselves?

This puzzle is similar to a well-known problem in the theory of public goods games. In such games, players are asked to contribute to a common pool. The content of this pool is then multiplied by a certain factor and divided equally among all players, irrespective of whether they contributed or not. If all players contribute, they all gain; but a selfish player who does not contribute gains even more. Consequently, if all players are selfish, there will be no common good. This is widely known as a social dilemma (or 'tragedy of the commons'). It is obvious that the defection of selfish players can be effectively prevented by punishing those who do not contribute. But if the punishment is costly to the punishers, this raises a 'second-order social dilemma' (see, e.g., Boyd and Richerson, 1992; Fehr and Gächter, 2000, 2002; Sigmund et al., 2001): indeed, a selfish player ought to refrain from punishing players who did not contribute, because this entails costs. Hence, no punishment, and hence, no public good. It may be argued that such a punishment can reform the punished player and turn him or her into a contributor for the future interactions. But as experiments by Fehr and Gächter (2002), see also Fehr and Fischbacher, 2003; have shown, the threat of punishment works even if players know that they are never going to meet with the punished co-player again.

Returning to indirect reciprocation, we meet a similar paradox with the scoring strategy. Refusing to help low-scorers is a way of punishing them: but this punishment is costly to punishers since it lowers their own score. Why should players incur such a cost if (as the model requires) they will never interact with the defector a second time? One possible escape from the dilemma is to assume that players who refuse to help a low-scorer (i.e., who 'justifiably' defect) will not have their own score reduced. This mechanism, which was suggested by Sugden (1986) already and briefly considered in Nowak and Sigmund (1998a), is usually termed the standing strategy. It is easily seen to be evolutionarily stable, see Leimar and Hammerstein (2001). But there are two problems with this solution. The one is that it begs the question: effectively, it means that punishment is not costly. Some game theorists feel that this solution to the social dilemma is (literally) too cheap to be interesting. The other objection is that such an accounting based on 'justifiable' defections requires considerable cognitive abilities from the players, and seems vulnerable to errors. Can it evolve under plausible conditions?

Experimental investigations by Milinski et al. (2001) indicate that in real-life human interactions, the mechanism based on standing is not as prevalent as the much simpler scoring mechanism. But here, we are interested in the theoretical aspects: for which parameter ranges (payoff values, error probabilities, etc.) must it be required that standing (rather than scoring) is the prevalent way of assessing actions by third parties, in order to see cooperation emerge?

We will therefore compare several possibilities for assessing the score of players. In the simplest case, the score depends entirely on how often the player has provided or refused help, irrespective of the moral standing of the co-players. A more sophisticated approach does not count the defections which are justified, i.e. addressed towards a co-player with a low score. And a third, rather stern way of judging players would be to reduce the score of those who do provide help to low-scoring co-players, because this lack of discrimination threatens to subvert the punishment system. To return to the parallel with the social dilemma with punishment, this would correspond to punishing, not only those who fail to contribute, but also those who fail to punish (see e.g. Boyd and Richerson, 1992; Gintis, 2000).

In addition, players who are in the position of the potential donor can use the score in several ways as a basis for their decision of whether or not to help. The simplest way, which we have tacitly assumed so far, consists in basing this decision entirely on the score of the potential recipient, and thus to help if and only if the recipient's score is sufficiently high. From the viewpoint of classical game theory, this seems bizarre (cf. Leimar and Hammerstein, 2001): if my future payoff depends only my own score, why should I base my decision on the co-player's score? From this viewpoint, it would be more sensible to base my decision on my own score, and give help if and only if my score threatens to become so low that my future chances of receiving help will be affected. In our simulations, decision rules based uniquely on the donor's own score do not lead to cooperative populations. But decision rules based on both the donor's and the recipient's score, or based exclusively on the recipient's score, can evolve and lead to cooperation.

We thus have a large set of potential strategies for indirect reciprocation, based (a) on different assessment rules for judging interactions between third parties (i.e. different ways in which giving or withholding aid affects the score of the potential donor, depending on the 'moral' status of the potential recipient) and (b) on different action rules (how to determine the decision of giving or withholding aid, based on the own score and

on that of the co-player). We study the evolution of such strategies, depending on the cost-to-benefit ratio, the probability of mistakes in implementation or perception, the average number of interactions within a lifetime, etc. The simulations can be performed online, see Brandt (2004).

Before launching into these investigations, two remarks are in order. The first concerns the structure of the population. In Nowak and Sigmund (1998a, b) and Panchanathan and Boyd (2003), Fishman et al. (2001) and Fishman (2003), populations are assumed to be well-mixed, typically consisting of 100 individuals each engaged in some five or ten interactions as a donor. The more elaborate statistics of Leimar and Hammerstein (2001) are based on a population structure likely to be a more realistic image of prehistoric mankind: 100 tribes of 100 players each, with some gene flow between the tribes. We shall adopt the Leimar–Hammerstein model, because it is apt to avoid spurious effects of random drift.

The second remark concerns the score range. Nowak and Sigmund introduced, besides their full model, a simplified version where the score takes only two values ('good' and 'bad') and where players remember only the last decision of their co-player. This allows to derive some analytical expressions for the payoff values, see Nowak and Sigmund (1998a, b), Panchanathan and Boyd (2003), Fishman (2003), and Ohtsuki (2004). In this context, Panchanathan and Boyd (2003) state flatly that 'indirect reciprocity cannot be based on an image-scoring strategy when errors are considered'. We claim that their analytical result concerning the instability of discrimination based on scoring need not hold for the original model by Nowak and Sigmund (1998a). For an alternative analytic approach based on the assumption that the number of rounds is Poisson-distributed, a stable mixture of discriminating and undiscriminating altruists can evolve.

2. Action and assessment modules

Every strategy in the present model consists of two modules, an assessment module and an action module. The assessment module comes into play when individuals observe interactions between two players. The image of the player acting as potential donor is possibly changed. The image of the recipient, who is the passive part in the interaction, remains unchanged. The action module prescribes whether a player in the position of a potential donor provides help or not.

Starting with the assessment module, we shall for simplicity assume that individual A's score of individual B depends only on how B behaved (towards some third party C) when last observed by A as a potential donor. Thus A has a very limited memory, and the score of B can only take two values, *good* and *bad*. We shall assume that all players are born *good*. In every interaction observed by A, there are two possible outcomes (B can give help or not), two possible score values for B and two for C. Thus there are eight possible types of interaction, and hence, depending on whether they find A's approval or not, $2^8 = 256$ different value systems.

In a first approach, based on common intuition, we will consider only three of these value systems, or 'morals'. We shall say that they are based on SCORING, STANDING and JUDGING, respectively (these terms are not completely felicitous, but the names of the first two, at least, are fixed by common use). These morals differ on which of the observed interactions incur reprobation, i.e. count as *bad*. Someone using the SCORING assessment system will always frown upon any potential donor who refuses to help a potential recipient, irrespective of the latter's image. Someone using the STANDING assessment system will condemn those who refuse to help a recipient with a *good* score, but will condone those who refuse to help a recipient with a *bad* score. Those using the JUDGING assessment system will, in addition, extend their reprobation to those who help a player with a *bad* score.

Thus these three value systems are of different strictness towards wrong-doers. Roughly speaking, someone who refuses to help is always bad in the eyes of a SCORING assessor. Only those who fail to give to a *good* player are bad in the eyes of a STANDING assessor. Someone who fails to give to a *good* player, but also someone who gives help to a *bad* player is bad in the eyes of a JUDGING assessor (see Table 1).

Turning to the action module, we shall assume that a player's decision on whether to help or not is based entirely on the scores of the two players involved (the own and that of the potential recipient), and not on other factors (i.e. the current payoff, the amount of help

Table 1
Assessment modules

Observed situation	SCORING	STANDING	JUDGING
G good/good	Good	Good	Good
G good/bad	Good	Good	Bad
G bad/good	Good	Good	Good
G bad/bad	Good	Good	Bad
N good/good	Bad	Bad	Bad
N good/bad	Bad	Good	Good
N bad/good	Bad	Bad	Bad
N bad/bad	Bad	Bad	Bad

According to the three assessment modules SCORING, STANDING, and JUDGING, individuals associate images to observed players chosen as potential donors, depending on the images of the potential donor/the recipient, and on whether a gift was given ("G"), or not ("N").

Table 2
Action modules

Donor/recipient	SELF	CO	AND	OR	AllC	AllD
Good/good	N	G	N	G	G	N
Good/bad	N	N	N	N	G	N
Bad/good	G	G	G	G	G	N
Bad/bad	G	N	N	G	G	N

Depending on their own image and the image of the recipient, individuals chosen as potential donor decide whether to give a gift ("G"), or not ("N"), according to their action module.

received so far...). Since there are four situations (donor and recipient can each be *good* or *bad*), there are $2^4 = 16$ possible decision rules, but we shall restrict attention to only four of them. CO is uniquely affected by the score of the potential recipient, and gives if and only if that score is *good*. SELF worries exclusively about the own score, and gives if and only if this score is *bad*. AND gives aid if the recipient's score is *good* and the own score *bad*, and OR gives aid if the recipient's score is *good* or the own score *bad*. In addition to these four types, we also consider the two unconditional strategies, always to give, and never to give, AllC and AllD, which do not rely on scores at all (see Table 2).

A strategy in this model for indirect reciprocity is determined by a specific combination of action and assessment strategy. Since the strategies AllC and AllD do not rely on scores, no assessment module has to be specified. Thus we compare 14 different strategies: the four action modules CO, SELF, AND and OR, each combined with the three assessment modules SCORING, STANDING and JUDGING, and the two unconditional strategies AllC and AllD. We will denote strategies which always entail cooperation between *good*-standing individuals as cooperative (CO, OR, and AllC) and the other strategies (SELF, AND and AllD) as defective.

We consider a population consisting of M tribes with N members each. Within each tribe, n interactions take place per generation. For each of these interactions, two individuals are randomly chosen within the tribe and assigned the roles of potential donor and potential recipient. Each individual will thus make on average $n/2N$ decisions on whether to help or not. In our simulations, this number will always be small.

If the potential donor decides to help, the recipient obtains a benefit b at a cost $-c$ for the donor. Following common practice (see Nowak and Sigmund, 1998a; Leimar and Hammerstein, 2001), we add the amount c in each interaction to both the potential donor and the potential recipient in order to avoid negative payoff values. We note that the results remain essentially unchanged if instead of adding the amount c in every interaction a certain baseline fitness is introduced. We normalize by setting $c = 1$.

The outcome of an interaction is observed by some (or all) co-players in the tribe. Each individual keeps a (private) score of all tribe-members. This observation could, in principle, be based on hearsay, but we shall not consider the strategy of spreading false rumours.

In the course of a generation, all individuals keep their strategy and accumulate payoff. After all n interactions have taken place, the offspring generation is formed by assuming, as in Leimar and Hammerstein (2001), that within each tribe, an offspring individual is derived locally with probability p and non-locally with probability $1 - p$. The locally derived individuals inherit the strategy of a tribe-member of the previous generation, the non-locally derived individual inherit the strategy from a member at large. In each case, the probability that an individual from the parent generation is chosen is proportional to that individual's payoff.

Within this model, we shall consider the influence of the following parameters: the average number of rounds, i.e. n/N; the number of social groups M and the migration rate $1 - p$; the benefit-to-cost ratio, i.e. the value of b; the initial composition of the population; the mutation rate (in particular, we will consider the influence of a steady influx of AllD players); the number of generations; and the error rates. We shall investigate both errors in implementation and errors in perception. We shall also assume that not all interactions are watched by all players, and thus assume, for instance, that only one, or two, or three out of four interactions are witnessed by a given individual.

We shall consider two distinct types of scenarios: full runs, with all 14 strategies initially present (and reintroduced in case mutations are allowed), and limited runs, where one module is specified (for instance, the assessment module STANDING or the action module CO) and the other module explores the three or four possible alternatives.

2.1. Agent-based simulations

In this section we test a basic model and do not allow for errors.

We consider a population of $M = 100$ tribes, each consisting of $N = 100$ individuals. The population is initialized randomly, all 14 strategies being equally probable. See Fig. 1a for the representation of these strategies.

In every generation, and in every group, $n = 1000$ games will be played. The resulting average number of rounds $n/N = 10$ ensures that practically every player is involved in a few interactions, both as donor and recipient, and has enough possibilities to update the co-players' images. If the number n of games is too small, strategies cannot display their characteristics. A large

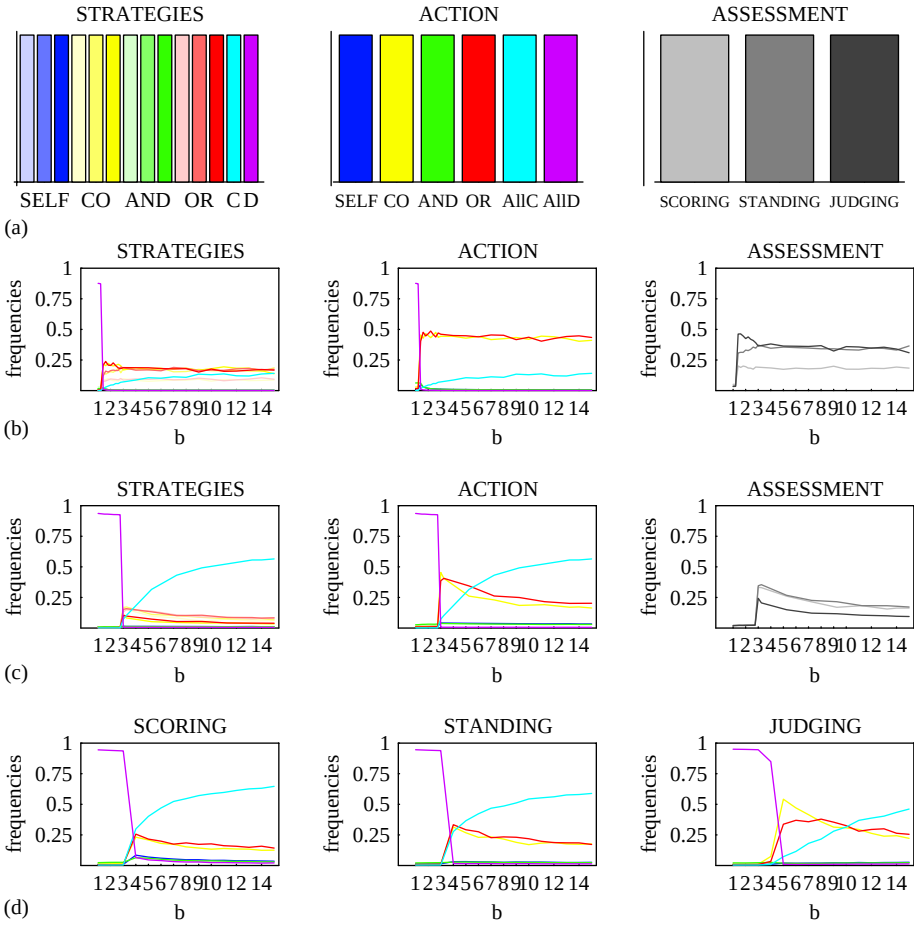

Fig. 1. (a) Color scheme for the following plots. Action modules are presented by different colors, whereas assessment modules are specified by their brightness. (b) Long-term frequencies of strategies of a population of $G = 100$ tribes, each with $N = 100$ individuals and $n = 1000$ games per generation. The first plot shows the frequencies of all 14 strategies, whereas the other plots show the frequencies of action modules and assessment modules, respectively. In the second plot, long-term frequencies for each action module are summed up over the three assessment modules, except for AllC and AllD which are shown at their actual frequencies. In the third plot, frequencies for each assessment module are summed up over all action modules, and AllC and AllD are not taken into account. For $b < 1.5$, AllD dominates, but for higher benefit-to-cost ratios, the cooperative strategies CO, OR, and AllC dominate and effectively suppress defectors. CO and OR perform nearly equally well, AllC reaches comparable frequencies only for high b. Parameters are $p = 0.9, \mu = 0.01$, errors are not included ($p_o = 1, \varepsilon_a = \varepsilon_o = 0$). Frequencies are averaged over generations 9000–10 000, averaged over 10 runs. (c) If errors in perception and action are included, defective strategies are suppressed to a minimum for $b \geqslant 3$, and high frequencies are reached by AllC. JUDGING is most vulnerable to errors, SCORING and STANDING attain practically equal levels. Parameters are $G = 100$, $N = 100$, $n = 1000$, $p = 0.9$, $p_o = 0.5$, $\varepsilon_a = \varepsilon_o = 0.05$, $\mu = 0.01$. Frequencies are obtained as above. (d) Comparison of the long-term frequencies for the assessment modules SCORING, STANDING, and JUDGING, from left to right. All types of errors and mutation are included. For SCORING and STANDING, a cooperative outcome is achieved for $b \geqslant 4$, compared to $b \geqslant 5$ for JUDGING. However, SCORING has to tolerate a small proportion of defective strategies even for high benefits. For $b \geqslant 4$, frequencies of cooperative acts are around 85 percent for SCORING and STANDING, whereas JUDGING attain 80 percent only for very high b.

number, on the other hand, does not significantly change the results.

We examine for which values of the benefit-to-cost ratio b, cooperative strategies become more frequent than defective strategies, and analyse the long-term frequencies of the strategies for the cooperative regime. For smaller values, AllD is invariably established as the most successful strategy within a few generation.

If there is no gene flow between the groups ($p = 1$), defective strategies vanish almost completely from the population within 100 generations for $b \geqslant 2$. CO and OR matched with STANDING and JUDGING are the most successful strategies and survive at nearly equal frequencies. AllC reaches comparable frequencies for higher b-values. CO and OR matched with SCORING each reach around 10 percent. Once there are only the action modules CO, OR and AllC left in the population, their strategies do equally well, since in this error-free scenario, all individuals keep their *good* image and cooperate. Around 98 percent of all interactions are cooperative. For $b = 1.5$, AllD survives at nearly 20 percent, whereas SELF and AND are suppressed to very small values.

For $b \geqslant 1.5$, even a small gene flow ($p = 0.99$, for instance) guarantees that in no tribe, defective strategies get established. The action modules CO and OR dominate the other strategies after several generations, as before. For $b \geqslant 2$ the defective strategies SELF, AND, and AllD usually vanish, and the frequency of cooperative acts reaches 100%. These results continue to hold for higher gene flow.

Next, we allow for mutations: with a probability μ, a new individual at the beginning of a generation is formed, not by drawing from the local or global gene pool, but by assigning it one of the 14 strategies with equal probability. No strategy can die out completely, and well-established strategies keep getting confronted with a minority of dissidents.

Fig. 1b presents a summary of the outcome in the long run for different b-values in the range $1 \leqslant b \leqslant 14$. The frequencies of strategies are averaged over 1000 generations after an initial phase of 9000 generations. Since for $b > 2$ a cooperative outcome is usually achieved within 100 generations, these statistics present reliable information about the long run. We used the high mutation rate of $\mu = 0.01$, so that strategies have a chance to reinvade. However, the simulation results do not significantly differ from the case of no mutation. Obviously, the established strategies are immune to invasion by mutants. For $1.5 < b < 2$, AllC individuals cannot survive, and only the CO and OR modules are present. These two action modules are responsible for suppressing minorities of defectors. If the benefit-to-cost ratio b is higher, AllC benefits from CO and OR individuals, who discriminate wrong-doers, and thus unconditional cooperators also get a firm foothold. For $b = 2$, some 95 percent of all interactions are cooperative, and for larger values of b help is refused in fewer than 2 percent of the interactions.

2.2. The role of errors

The decisions of the players depend on their information. If only part of the interactions between third parties are observed, the performance of the strategies will be altered. Moreover, errors in both the action and the assessment module are likely to have strategic consequences. In the following section we investigate the role of different types of errors, and the corresponding robustness of the strategies.

2.2.1. Not all games are observed

Let p_o be the probability that an individual observes a given interaction between two other members of the tribe. If on average only every second interaction is observed by the players ($p_o = 0.5$), then a cooperative outcome is still achieved, usually within 100 generations, for $b \geqslant 2.5$. For $b \geqslant 3$, defective strategies are reduced to a minority of around 2 percent in case mutations are allowed. If we exclude mutation, defective strategies die out completely. CO and OR combined with SCORING and STANDING survive at nearly equal frequencies (some 12–14 percent); combined with JUDGING, these two strategies attain 6 to 8 percent. Profiting from their success, AllC individuals, whose strategy is unaffected by $p_o > 0$, outcompete the other cooperative strategies: their long term frequencies range at 30 percent for high values of b.

JUDGING is affected most severely from the fact that not all interactions can be observed: for $b = 5$, JUDGING individuals associate wrong images to 15 percent of their co-players, whereas SCORING and STANDING err for only 3 percent. JUDGING individuals are therefore less frequent in the population (only 15–20 percent, compared to 25–30 percent for SCORING and 30–35 for STANDING). This decline in frequency, however, holds only when mutations are excluded. For $\mu > 0$, *JUDGING* individuals can spread as soon as defectors are eliminated, and the three assessment modules reach almost the same long term frequencies.

It is clear that if fewer games are observed, it becomes more difficult for cooperative strategies to survive. Whereas for $p_o = 0.25$ a cooperative outcome is achieved for $b \geqslant 3$, for $p_o = 0.1$ the threshold is shifted to $b \approx 6$. Again, differences in the assessment modules become apparent: JUDGING individuals are more likely to get wrong images of their coplayers (20 percent, versus 5 percent for SCORING and STANDING players).

2.2.2. Errors in implementation

Let ε_a be the probability of an error in the implementation of a move. An individual in the role of a donor then performs the opposite of the intended action. This can occur, in particular, if an individual happens not to have the resources for helping the co-player (cf. Fishman, 2003).

In the absence of mutation and other types of errors, errors in implementation have surprising effects: For $b \geqslant 2$, JUDGING reaches frequencies of more than 95 percent, leaving a small minority of STANDING individuals. Thus practically the whole population consists of CO and OR players with the JUDGING module. AllC vanishes completely. A steady flow of mutations allows STANDING and SCORING modules to survive at higher frequencies, but JUDGING still reaches frequencies of around 60 percent for $b \geqslant 2$. Therefore JUDGING seems, at first sight, robust against errors in implementation. However, this outstanding success of JUDGING requires that every situation is observed by all members of the population.

The situation is altered if not all games are watched by coplayers. For $p_o = 0.5$, i.e. if on average only every second interaction is observed, JUDGING individuals

can no longer successfully rely on their images. This is due to the fact that JUDGING is poor at correcting wrong images. In the absence of mutations, the JUDGING module vanishes almost entirely from the population. Nevertheless, the STANDING module and (to a lesser degree) the SCORING module ensure that almost full cooperation is established for $b \geqslant 3$, with a high frequency of AllC players. With mutation, however, the JUDGING module can subsist, although at lower frequencies than the other two assessment modules.

2.2.3. Errors in perception

Let us consider now errors in perception and assume that with a small probability ε_o, an individual misunderstands a situation, and assigns the wrong image to a player. If we set $\varepsilon_o = 0.05$ and exclude mutation and other types of errors, the simulation results are quite different from the case of errors in implementation: JUDGING individuals vanish completely from the population. STANDING is more successful than SCORING. For $b \geqslant 2$, cooperation dominates, and AllC becomes very frequent.

If mutations are allowed, JUDGING can re-invade the population, but will still form the least frequent assessment module. Approximately 30 percent of the images of JUDGING players are incorrect, and only 5 percent for the two other assessment modules. Cooperation occurs in 90 percent of all interactions.

In contrast to the case of errors in implementation, a reduced probability of observing interactions now has little additional effect. AllC reaches high frequencies, and the JUDGING module practically vanishes. SCORING and STANDING perform on a similar level. For $p_o = 0.5$ and $b \geqslant 2.75$ nearly full cooperation is achieved.

2.2.4. Combination of all types of errors

A combination of all types of errors together with possible mutations confirms that CO, OR and AllC dominate for $b \geqslant 3$, see Fig. 1c.

Due to the high percentage of wrong images (around 46 percent), JUDGING is the least successful assessment module, whereas SCORING and STANDING perform nearly equally well. Even when the chosen mutation rate as well as the error rates are high, more than 80 percent of all games are cooperative for $b \geqslant 4$.

2.3. Comparison of assessment modules

In this section we test separately each of the three assessment modules. We include all types of errors and mutations, and allow for only six strategies in each simulation, namely all four action modules combined with the fixed assessment module, and the two unconditional strategies AllC and AllD. All six strategies are initially equally frequent, so that AllC and AllD start with higher frequencies than in the previous simulations, see Fig. 1d.

For SCORING and STANDING, cooperation dominates for $b \geqslant 4$. However, when using the SCORING modules, the population has to tolerate a minority of ≈ 6 percent of AllD, whereas STANDING is able to suppress AllD to around 1.5 percent. For high benefits b, more than 85 percent of all games are cooperative (slightly more for STANDING than for SCORING).

The JUDGING module can establish cooperation only for $b \geqslant 5$, and the frequency of helpful acts reaches only 80 percent even for very high b.

The highest frequency of AllC players occurs for SCORING. This makes it more difficult to totally suppress defection. Defectors can therefore re-invade a SCORING regime more easily than a STANDING regime. JUDGING, which is the strictest module, reduces AllC to quite low frequencies for intermediate values of b, because AllC individuals are seen as *bad* after giving to a *bad* individual. But the JUDGING population cannot achieve the high level of cooperation achieved by the two other modules.

Irrespective of the assessment module, the cooperative strategies CO, OR and AllC dominate for high benefit-to-cost ratios. Obviously, AllC can only thrive in the presence of more discriminating cooperative strategies. In order to test whether CO or OR are more effective in promoting cooperation, we take one of them out of the game and see how the remaining action modules perform. We do this separately for each of the three assessment modules. It turns out that STANDING works better with OR than with CO, in the sense that cooperation gets established for $b \geqslant 4$ if CO is deleted, and only for $b \geqslant 5$ if OR is deleted. In the case of SCORING, the situation is the opposite: if OR is excluded, cooperation dominates for $b \geqslant 5$, while if CO is excluded, cooperation emerges only for $b \geqslant 6$. For JUDGING the difference is even more pronounced: without OR, cooperation is established for $b \geqslant 5$, as compared to $b \geqslant 7$ without CO. The OR module combined with STANDING does best.

2.4. Separate performances

In this section we investigate how good each strategy is at invading a majority of defectors, by assuming that the frequency of the invading strategy is 10 percent and all other players use AllD. We do not allow for errors or mutation.

SELF cannot invade, and vanishes completely, irrespective of the assessment module and the benefit-to-cost ratio. CO combined with STANDING or JUDGING takes over the whole population for $b \geqslant 4.5$, whereas when combined with SCORING, CO can only invade for $b \geqslant 6.5$. AND suppresses defectors

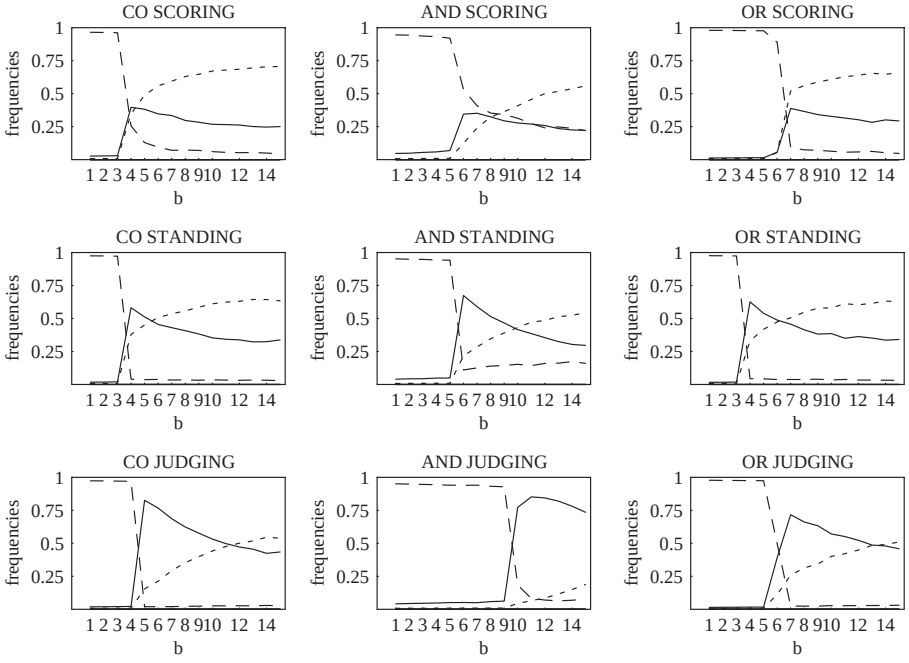

Fig. 2. In each plot, long-term frequencies for only three strategies are shown: AllC (short dashes), AllD (long dashes), and an additional discriminating strategy (solid line). SELF cannot promote cooperation and is therefore not shown. For all action modules, STANDING performs best, although for CO its advantage is small. Parameters are $G = 100$, $N = 100$, $n = 1000$, $p = 0.9$, $p_o = 0.5$, $\varepsilon_a = \varepsilon_o = 0.05$, $\mu = 0.01$. All three strategies are equally frequent initially.

only if $b \geqslant 10$, doing best when combined with STANDING. OR combined with STANDING or JUDGING takes over the population for $b \geqslant 4.5$, whereas OR combined with SCORING works only for $b \geqslant 11$.

The presence of AllC players makes it more difficult to suppress defectors. We therefore analyse populations with AllC, AllD, and one third strategy initially present, in equal proportions, see Fig. 2. All types of errors are included. Again, SELF cannot promote cooperation, irrespective of the assessment module, and AllD takes over practically the whole population. Even for high b-values, SELF forms at best a minority of 3 percent, and AllC levels at 1 percent. The frequency of cooperative acts falls below 8 percent within one generation.

The CO modules works best with STANDING, and defectors vanish for $b \geqslant 3.5$, compared to $b \geqslant 4.5$ for the two other assessment modules. SCORING and STANDING reach levels of cooperation around 90 percent, whereas in the case of JUDGING, only ≈ 80 percent of all interactions are cooperative. Moreover, JUDGING has to tolerate a high frequency of wrong images (≈ 45 percent, as compared to ≈ 12 percent for the other two assessment modules). In the case of SCORING and STANDING, and for high b-values, AllC takes advantage from the fact that CO suppresses defectors, and even outcompetes CO.

Quite surprisingly, the defective action module AND promotes AllC players and cooperative acts occur with frequencies of up to 60 percent if b is large. Although AND-individuals only give if they have a *bad* image of themselves and are matched with a coplayer seen as *good*, this strategy is able to promote cooperation. Despite the fact that the AND-module differs only in one situation from the SELF-module, and refuses help in this situation, the outcome for the two strategies is quite unrelated. Once more, STANDING is the most successful assessment module, but it has to tolerate up to 10 percent of AllD players.

The action module OR reveals large differences for the assessment modules. OR combined with STANDING can suppress defectors for $b \geqslant 3.5$, and reaches 88% of cooperative acts. With SCORING and JUDGING, OR can suppress AllD only for $b \geqslant 7$ and $b \geqslant 6$, respectively.

All simulation results reveal that the CO and OR strategy are most successful in promoting a cooperative outcome.

2.5. *SCORING with errors*

Analytical expressions for the payoff values are easiest to derive, except for the SCORING module. If we consider a well-mixed population consisting of only of AllC, AllD and (as unique discriminating strategy) CO-SCORING, and if we allow, for simplicity, only implementation errors turning an intended cooperation into a defection with a certain probability $1 - r$, we can calculate the payoff values (see Nowak and Sigmund (1998a, b), Panchanathan and Boyd, (2003) and Fishman (2003)). Let us denote the frequencies of the three

strategies by x, y and z, respectively, and by $P_i(m)$ the payoff values for the three strategies in the mth round, for $i = 1, 2, 3$. Since the replicator dynamics is unchanged by adding the same function to each of these values, we can normalize by setting the defectors' payoff equal to zero. This yields $P_2(m) = 0$ (trivially), and for $m > 1$:

$$P_1(m) = -r + br^2 z,$$

$$P_3(m) = r(b-1)G_m - br^2 x,$$

where G_m, the frequency of players with a *good* image in the mth round, is given by

$$G_m = \frac{rx[1-(rz)^{m-1}]}{1-rz} + (rz)^{m-1}$$

(clearly $G_1 = 1$ and $P_1(1) = P_3(1) = -r$).

Panchanathan and Boyd (2003) have shown that in the presence of errors (i.e. for $r < 1$), if one assumes that there exists a constant probability $w < 1$ for a further round, there exists a unique equilibrium with $y = 0$ (no defectors) involving a mixture of discriminating and undiscriminating altruists. This equilibrium is unstable, however, and the replicator dynamics leads in the long run to $y = 1$. Panchanathan and Boyd conclude that indirect reciprocity cannot be based on an image scoring strategy when errors are included. This is in contrast to the numerical simulations obtained here.

The reason for the discrepancy lies in the fact that Panchanathan and Boyd assume a geometric distribution for the number of rounds for each player. This is in line with the usual assumptions about the Prisoner's Dilemma game. But in the simulations presented here (as well as in Nowak and Sigmund (1998a, b) and Leimar and Hammerstein (2001)), the number of rounds is given by a binomial distribution, corresponding to a Bernoulli scheme with n trials and a success probability of $1/2N$ (the player has to be sampled, and assigned the role of the donor). Let us approximate this by a Poisson distribution with parameter $\lambda = n/2N$, so that the probability that the player is engaged in exactly k rounds ($k = 0, 1, ...$) is given by

$$w(k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!},$$

where λ is the average number of rounds.

We shall show that for $b > 2$, there always exists a stable mixture of discriminating and undiscriminating altruists provided λ, the average number of rounds, is sufficiently large. (For $b = 3$, for instance, it suffices that there are more than 2 rounds on average.) See Fig. 3. Thus SCORING can allow cooperation based on indirect reciprocity to evolve.

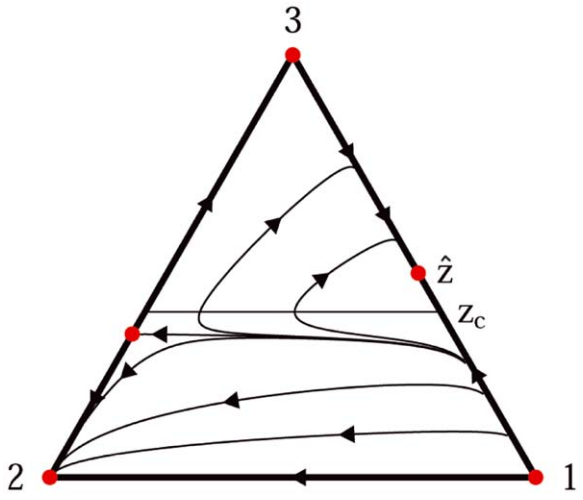

Fig. 3. Phase portrait for the replicator dynamics with the strategies AllC, AllD and CO SCORING. A mixture of the two cooperative strategies is stable. Parameter values are $b = 5, \lambda = 2$, and $r = 0.9$.

The average payoff for a player using strategy i is given by

$$P_i = w(0)0 + w(1)P_i(1) + w(2)[P_i(1) + P_i(2)] + \cdots = \sum_{k=1}^{\infty} w(k)\left[\sum_{m=1}^{k} P_i(m)\right].$$

Up to a factor $r > 0$ which we shall henceforth omit, we obtain for the expected total payoff values P_i of the three strategies, $P_2 = 0$ (obviously),

$$P_1 = brz(\lambda - 1 + \mathrm{e}^{-\lambda}) - \lambda$$

and

$$P_3 = -brx(\lambda - 1 + \mathrm{e}^{-\lambda}) - b + b\mathrm{e}^{-\lambda} + (b-1)\frac{rx\lambda}{1-rz} + \frac{(b-1)(1-rx-rz)}{(1-rz)^2}[1 - \mathrm{e}^{\lambda(rz-1)}].$$

In these expressions, z and r occur only within the product zr.

Setting $s := 1 - rz$ and using the function $f(u) = u - 1 + \mathrm{e}^{-u}$ (a function which is positive, strictly increasing and convex for $u \geqslant 0$ and satisfies $f(0) = 0$), we get

$$P_1 = (1-s)bf(\lambda) - \lambda.$$

On the edge $y = 0$ we have

$$P_3 - P_1 = (1-r)g(s),$$

with

$$g(s) = bf(\lambda) - (b-1)\frac{f(\lambda s)}{s^2}.$$

We note that $g(1) = f(\lambda) > 0$ and

$$g(0) := \lim_{s \downarrow 0} g(s) = bf(\lambda) - (b-1)\frac{\lambda^2}{2}$$

which is negative for sufficiently large λ. Furthermore, g is monotonically increasing. Hence there exists a unique $\hat{s} \in]0,1[$ such that $P_1(\hat{s}) = P_3(\hat{s})$. This defines a $\hat{z} = (1-\hat{s})/r$ and hence an equilibrium on the edge $y = 0$, if r is sufficiently close to 1. (Indeed, on that edge, $\dot{z} = zx(P_3 - P_1)$). In order to prove the stability of the corresponding fixed point in the (x, y, z)-simplex, all that remains to be shown is that $P_1(\hat{s}) > 0$ or equivalently that $\hat{s} < s_c$, where s_c is defined by $P_1(s_c) = 0$ (see Fig. 4).

This in turn follows from the fact that the function g is concave, and that the intercept s^* of the chord joining $(0, g(0))$ with $(1, g(1))$ lies to the left of s_c. Now $s^* < s_c$ is equivalent to $b > \alpha + \alpha^2$, with $\alpha := f(\lambda)/\lambda$. If $b > 2$, this holds for sufficiently large λ, since $\alpha \to 1$ for $\lambda \to +\infty$.

2.6. Discussion

Nowak and Sigmund considered all four types of strategies based on SCORING. They stressed that 'a discriminator who punishes low-score players by refusing them help pays for this by having his own score reduced' (Nowak and Sigmund, 1998ba), and that 'the success of a discriminating player is somewhat hampered by the fact that whenever he refuses to help a 'bad' scorer, he loses his 'good' score' (Nowak and Sigmund, 1998bb). This aspect was also stressed by Ferrière (1998). Nowak and Sigmund also wrote that STANDING strategies seemed intuitively plausible (but difficult to analyse), and stated: 'we believe that Sugden's strategy is a good approximation to how indirect reciprocity actually works in human communities' (Nowak and Sigmund, 1998bb).

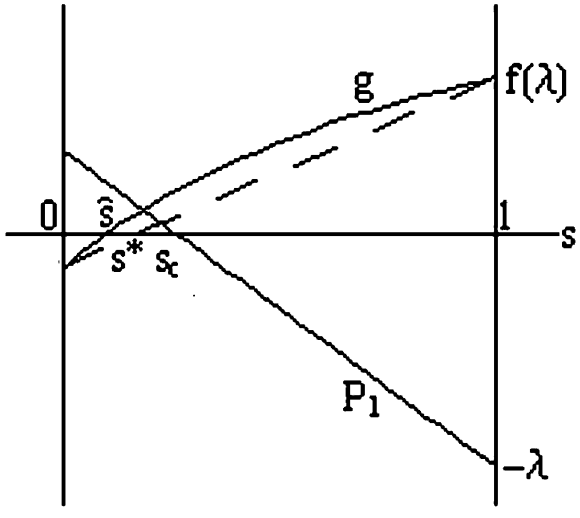

Fig. 4. Graph of the functions P_1 and g as described in Section 2.5. The value of $\hat{s}$ is to the left of s_c. Parameter values are $b = 2$ and $\lambda = 3$.

Leimar and Hammerstein (2001) compared the CO strategies for SCORING and STANDING, and found that STANDING was more robust. Panchanathan and Boyd (2002) considered several strategies: their DISC corresponds to SCORING-CO, RDISC is STANDING-CO and CTFT is STANDING-OR. Their basic message is that it is not crucial that individuals attend to their own standing: what is crucial is whether they are able to discriminate between justified and unjustified defection. Panchanathan and Boyd assume that all players agree on the images of all co-players, due to the effect of gossip. But this raises the possibility of deception. We assumed, in contrast, that the image of a given player is in the eye of the beholder, i.e. the same player can have different images in the eyes of different individuals.

Our results confirm that STANDING is almost in any respect superior to SCORING. This is not surprising, and corresponds to the obvious fact that cooperation based on punishment is easier to implement if punishment is non-costly. What we would like to stress here is that just as costly punishment is a realistic assumption which promotes cooperation, so SCORING is sufficient to uphold indirect reciprocation, even in the presence of errors.

In the experiments of Milinski et al. (2001), SCORING was more prevalent than STANDING. Most strikingly, they found that people who justifiably defect compensate for this by being more generous in the following rounds, apparently feeling the need to make up for a loss in their own score. We thus cannot follow the argument of Panchanathan and Boyd (2003) who claim that the interpretation of Milinski et al. (2003) is misleading.

This being said, all models so far are extreme simplifications of what is likely to occur in real human societies. Fishman (2003), for instance, makes the point that players will occasionally be unable to provide help, even if they wish to do so (the 'phenotypic defectors' of Lotem et al., 1999). It seems plausible that strategies for indirect reciprocation depend on the distribution of scores in the population, on the likelihood of future rounds, on the accumulated payoff, etc.

In particular, it must be stressed that the reduction to two score-values only ('good' and 'bad') makes the emergence of cooperation more difficult (but helps in keeping the number of possible strategies fairly low). Thus the SELF rule is reduced to helping every second time. With a larger range of score values, many simple strategies are less vulnerable to errors. A misunderstanding can lead to a score-value of 0, but in the next interaction as a donor, this can be offset and the score is as good as before. Furthermore, the artificial reduction to two score-values only, while allowing for some

analytical investigations, is likely to introduce spurious effects.

To give an example, Panchanathan and Boyd (2003) analysed the dynamics of the CO-SCORING strategy if errors in implementation are allowed. They concluded that the mixture of discriminating and indiscriminating altruists is unstable, and that defection will be the only long-term outcome. Fishman (2003) studied the dynamics under the assumption that players did, occasionally, not have the resources to actually provide help, even if they wished. He found that a mixture of discriminating and indiscriminating altruists is stable. Both models actually analyse similar situations, so that the different outcomes seem to contradict each other. But Fishman assumed a fixed number of interactions per individual, whereas Panchanathan and Boyd assumed, instead, a fixed probability w for a further round, so that the number of interactions was a random variable with expected value $1/(1-w)$. We have seen that the assumption of having a Poisson-distribution for the length of an individuals' 'lifetime' leads to a stable mixture of discriminating and undiscriminating altruists. Thus the outcome depends sensitively on modelling assumptions concerning the probability distribution of the number of rounds in the game.

More significantly perhaps, the two-score range is an artificial restriction which hampers the evolution of indirect reciprocity because it is more vulnerable to errors. Other simulations, which will be presented in a sequel to this paper, indicate that with a full range of score-values, some outcomes are different.

To resume, it is usually better to distinguish between justified and unjustified defections, just as it is better to avail oneself of non-costly rather than costly forms of punishment. But even if such a distinction is not available, discrimination can lead to robust regimes of overall cooperation in interactions based on indirect reciprocity.

Added in proof

We submitted this paper without being aware of the parallel work of Ohtsuki and Iwasa (published in this issue). In spite of the similarity of our approaches, it seems worth to publish both papers 'back to back', without modifying more than the terminology, because they lead to different (but by no means contradictory) conclusions. Ohtsuki and Iwasa look for those strategies which are evolutionarily stable (or more precisely, strict Nash equilibria, and therefore necessarily pure). Such pure strategies cannot be invaded, once established. In our approach, we were more interested in the process of invasion and, more generally, the underlying replicator dynamics, which often leads to polymorphic outcomes. Such polymorphisms can lead to the stable establishment of overall cooperation even if the discriminating strategy does not become fixed. Two more technical points: in Ohtsuki and Iwasa's model, individuals experience infinitely many interactions. This allows to derive analytical expressions and achieve remarkably general conclusions. We assumed, in contrast, very few interactions per lifetime. This is certainly not a necessary assumption: but if we have in mind evolution in a small-scale society, a great many interactions would necessarily mean repeated interactions with the same co-players, and hence opportunities for direct reciprocation. The other technical difference is that the image of a given player is the same for all in Ohtsuki-Iwasa's model, whereas we assumed private images. Finally, Ohtsuki and Iwasa assumed that the 'moral' was fixed in the population. The possibility of 'mutant' morals (i.e. alternative assessment modules) gives plenty of scope for neutral drift to lead to polymorphic populations. It could well be that the assumption of a binary image (only 'good' and 'bad') is too limited to deal with the evolution of morals in the context of indirect reciprocity.

Acknowledgements

H. Brandt and K. Sigmund acknowledge support from the Wissenschaftskolleg WK W008 "Differential Equations" of the Austrian Science Fund FWF.

References

Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter, New York.

Axelrod, R., Hamilton, W.D., 1981. The evolution of cooperation. Science 211, 1390–1396.

Bolton, G., Katok, K., Ockenfels, A., 2001, Indirect reciprocity in an Image Scoring Game. Working Paper.

Boyd, R., Richerson, P.J., 1992. Punishment allows the evolution of cooperation (or anything else) in sizeable groups. Ethol. Sociobiol. 113, 171–195.

Brandt, H., 2004. URL: http://mailbox.univie.ac.at/hannelore.brandt/a&a Interactive java applets for online simulations on action and assessment modules

Camerer, C.E., 2003. Behavioral Game Theory. Princeton University Press, Princeton, NJ.

Fehr, E., Fischbacher, U., 2003. The nature of human altruism. Nature 425, 785–791.

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90, 980.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137.

Ferrière, R., 1998. Help and you shall be helped. Nature 393, 517–519.

Fishman, M.A., 2003. Indirect reciprocity among imperfect individuals. J. Theor. Biol. 225, 285–292.

Fishman, M.A., Lotem, A., Stone, L., 2001. Heterogeneity stabilises reciprocal altruism interaction. J. Theor. Biol. 209, 87–95.

Gintis, H., 2000. Game Theory Evolving. Princeton University Press, Princeton, NJ.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocation. Proc. R. Soc. Lond. B 268, 745–753.

Lotem, A., Fishman, M.A., Stone, L., 1999. Poor phenotypes stabilize indirect reciprocity by image scoring. Nature 400, 226–227.

Milinski, M., Semmann, D., Bakker, T.C.M., Krambeck, H.J., 2001. Cooperation through indirect reciprocity image scoring or standing strategy? Proc. R. Soc. Lond. B 268, 2495–2501.

Milinski, M., Semmann, D., Krambeck, H.J., 2002. Donors in charity gain in both indirect reciprocity and political reputation. Proc. R. Lond. Soc. B 269, 881–883.

Nowak, M.A., Sigmund, K., 1998a. Evolution of indirect reciprocity by image scoring. Nature 282, 462–466.

Nowak, M.A., Sigmund, K., 1998b. The dynamics of indirect reciprocity. J. Theor. Biol. 194, 561–574.

Ohtsuki, H., 2004. Reactive strategies in indirect reciprocity. J. Theor. Biol. 227, 299–314.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness? reputation dynamics in indirect reciprocity. J. Theor. Biol., to appear.

Panchanathan, K., Boyd, R., 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J. Theor. Biol. 224, 115–126.

Seinen, I., Schram, A., 2001. Social status and group norms: indirect reciprocity in a helping experiment. Working Paper, University of Amsterdam.

Sigmund, K., Hauert, C., Nowak, M.A., 2001. Reward and punishment. Proc. Natl Acad. Sci. 98, 10757–10763.

Sugden, R., 1986. The Economics of Rights, Cooperation and Welfare. Basil Blackwell, Oxford.

Trivers, R., 1971. The evolution of reciprocal altruism. Quart. Rev. Biol. 46, 35–57.

Wedekind, C., Braithwaite, V.A., 2002. The long-term benefits of human generosity in indirect reciprocity. Curr. Biol. 12, 1012–1015.

Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.