THE ROYAL SOCIETY

doi 10.1098/rspb.2000.1573

Evolution of cooperation through indirect reciprocity

Olof Leimar[1]* **and Peter Hammerstein**[2]

[1]*Department of Zoology, Stockholm University, S-106 91 Stockholm, Sweden*
[2]*Innovationskolleg Theoretische Biologie, Humboldt-Universität zu Berlin, Invalidenstrasse 43, D-10115 Berlin, Germany*

How can cooperation through indirect reciprocity evolve and what would it be like? This problem has previously been studied by simulating evolution in a small group of interacting individuals, assuming no gene flow between groups. In these simulations, certain 'image scoring' strategies were found to be the most successful. However, analytical arguments show that it would not be in an individual's interest to use these strategies. Starting with this puzzle, we investigate indirect reciprocity in simulations based on an island model. This has an advantage in that the role of genetic drift can be examined. Our results show that the image scoring strategies depend on very strong drift or a very small cost of giving help. As soon as these factors are absent, selection eliminates image scoring. We also consider other possibilities for the evolution of indirect reciprocity. In particular, we find that the strategy of aiming for 'good standing' has superior properties. It can be an evolutionarily stable strategy and, even if not, it usually beats image scoring. Furthermore, by introducing quality variation among individuals into the model, we show that the standing strategy can be quality revealing, adding a new dimension to indirect reciprocity. Finally, we discuss general problems with currently popular modelling styles.

Keywords: evolution of cooperation; indirect reciprocity; image scoring; good standing

1. INTRODUCTION

Trivers (1971) suggested that the evolution of cooperative behaviour could be understood in terms of reciprocal aid giving in repeated interactions between two partners. This may be referred to as direct reciprocity whereas in indirect reciprocity the return from a social investment is expected from someone other than the recipient of the aid. According to Alexander (1979, 1987), indirect reciprocity is an important form of cooperation in complex human societies. Nowak & Sigmund (1998*a*) used computer simulations to show that indirect reciprocity could be implemented by a mechanism they called image scoring. The way the mechanism works is that an individual's score increases on every occasion he or she donates aid to a recipient and decreases when there is an opportunity to help someone in need but no help is offered. Image scoring is thought to take place in a social group of moderate size, where group members could keep track of everybody's score. The group stays together for a period of time during which individuals typically have several opportunities for giving and receiving help. Strategies based on image scoring are of the kind where one gives help only to those whose score is above a certain threshold. Nowak & Sigmund (1998*a*) performed extensive computer simulations and concluded that a cooperative regime based on image scoring can be evolutionarily stable. They also stressed the importance of their results in the understanding of the evolution of human social behaviour.

Here, we critically examine the Nowak & Sigmund (1998*a*) model. Our conclusion is that, while indirect reciprocity could evolve in principle, there are serious problems with the image-scoring strategies. The main weakness of these strategies lies in their failure to represent the true strategic interests of an individual. To see this, one only needs to consider whether an individual could ever benefit by basing decisions partly or wholly on the score of a potential recipient of help. In the setting described by Nowak & Sigmund (1998*a*), there seems to be no possibility for such a benefit. The only influence of an individual's current aid-giving decision on the probability of receiving aid in the future is due to the change in the individual's own score. A rational individual in this setting should then use a strategy that takes his or her own score into account, but ignores the score of a potential recipient. In addition, it would be in the individual's interest to know how other group members react to recipients with different scores in order to assess the consequences of a change in one's own score.

Our argument casts serious doubt on the evolutionary stability of strategies using the recipient's score as a basis for decision. In order to investigate the validity of our objection further we performed evolutionary simulations similar to those that seemed to support the image-scoring idea. We found that the previous success of image-scoring strategies was the result of restrictive conditions that are unlikely to be representative of historical human societies. Typically, image scoring would not evolve in a more realistic version of the scenario presented by Nowak & Sigmund (1998*a*).

In what appears to have been the first explicit specification of a strategy of indirect reciprocity, Sugden (1986) used the concept of good standing, which bears some similarity to the idea of an image score. In Sugden's model, everyone is initially in good standing. An individual loses good standing by failing to help a recipient in good standing whereas failing to help recipients who lack good standing does not damage the standing of a potential donor. An individual lacking good standing can also regain it by offering help when in position to do so. Sugden's 'standing strategy' is then to offer help when not in good standing or when the potential recipient is in good standing, otherwise a potential donor should not offer help. A crucial difference between this strategy and

*Author for correspondence (olof.leimar@zool.su.se).

Received 21 September 2000 *Accepted* 22 December 2000

the image-scoring strategies is that, in a population playing the standing strategy, it is in an individual's interest to react to the standing of a potential recipient. From Sugden's arguments it is clear that the standing strategy is evolutionarily stable. In our simulations we found that the strategy has additional robustness properties and that it can invade a population of image scorers. Thus, the standing strategy appears to be a viable candidate mechanism for human cooperation based on indirect reciprocity. Nevertheless, as we will discuss, there may well be other such candidates. In particular, there may be strategies where the role of good standing could be paralleled by aspects of individual quality that are revealed in acts of aid giving. Quite possibly, such strategies were on Alexander's (1979) mind when he developed his ideas of indirect reciprocity.

2. A NARROW SCOPE FOR IMAGE SCORING

(a) *The model*

Nowak & Sigmund (1998*a*) analysed a population consisting of a single social group of a size that could be representative of historical human societies, perhaps 100 or less. However, letting the entire population consist of a single social group has the consequence of introducing rather strong effects of random genetic drift. It is known that spatial genetic differentiation of human populations is relatively small, with Wright's F_{ST} varying from *ca.* 0.01 up to 0.07 for populations covering fairly large geographical regions (Cavalli-Sforza *et al.* 1996). This suggests that genetic drift has had rather small effects during human evolution. Thus, in order to study the case of moderately sized social groups while limiting the effects of genetic drift, we used the idealization of an island model (Wright 1943) in our simulations. In this model a certain proportion of the gametes forming a new generation are locally derived and the remainder are randomly drawn from the global gene pool. For real human populations, gene flow is more geographically restricted than in the island model. A formulation such as Kimura's stepping-stone model (Kimura & Weiss 1964) could also be considered and would be likely to produce results qualitatively similar to those for the island model. The crucial issue is to have a population structure such that the influence of genetic drift is limited.

Now consider a total population consisting of g social groups each with n group members, resulting in a total population size of $N=gn$. For the case $g=1$ our model reduces to that of Nowak & Sigmund (1998*a*). Interactions in a group consist of m rounds per generation. Two individuals are randomly chosen from the group in each round of interaction, one as a potential donor and the other as a potential recipient. If helping takes place, the cost c is subtracted from the donor's pay-off and the benefit b is added to the recipient's pay-off. In order to avoid negative pay-offs, we also add the amount c in each round to both the donor and recipient, as did Nowak & Sigmund (1998*a*). At the start of a generation, all group members have pay-off u_0, which can be zero or positive.

Strictly speaking, having a fixed number of m rounds rather than a stochastic number with expectation equal to m introduces the problem of the well-known 'end effect' (Luce & Raiffa 1957). With fixed m there can be no incentive to help in the last round and, therefore, no incentive to help in the next-to-last round, etc. However, since we will not introduce strategies that use the round number as a basis for decision in our simulations, we follow Nowak & Sigmund (1998*a*) in assuming a fixed number of rounds. Our argument does not depend on this assumption.

An individual's genotype specifies the strategy to be used during his or her lifetime. A new generation is formed by asexual reproduction as follows. One sums the pay-offs from the previous generation for each particular genotype for each group and normalizes to produce the within-group expected relative reproductive success. Similarly, summing over the entire population and normalizing gives the global expected relative reproductive success. A new individual is locally derived with probability p and is derived from the global gene pool with probability $1-p$. The genotype for each such individual is then randomly determined using a distribution corresponding to the appropriate relative reproductive success, local or global, provided there is no mutation. With a probability μ there is a mutation, in which case the genotype is equally likely to be any of the available genotypes. The algorithm implies that more productive groups make a greater contribution to the global gene pool but that each group contributes the same expected local proportion p through non-migrant individuals.

Each individual is endowed with a score s, which is set to zero at the start of a generation. A potential donor's score increases by one unit in a round of interaction if help is given and otherwise decreases by one unit (however, scores are constrained to stay within the range -5 to $+5$). An individual's score is known by all group members, for instance because all interactions are publicly observed. Given this information structure, a strategy could in principle base decisions to give help on the total set of past and present scores of the group members. It will be useful to consider a few special classes of such image-scoring strategies. They differ in how they use information about scores. One class uses only the donor's own score for decisions about aid giving, another only the recipient's score and yet another uses both scores. We will consider strategies of the form 'offer help when own score is less than the threshold h' for the first class in our simulations, the form 'offer help when recipient's score is at least k' for the second class and the form 'offer help when own score is less than h and recipient's score is at least k' for the third class, where h and k can vary. In addition, we will also introduce a class of strategies that were not studied by Nowak & Sigmund (1998*a*) in which decisions are based on one's own score and on the outcome of previous rounds of interaction between group members.

It is often the case in models of cooperation that differences between strategies only become apparent in a situation where some perturbation is explicitly introduced. The approach most commonly used to implement such a perturbation is to assume a probability of error in the execution of strategies (Selten & Hammerstein 1984; Sugden 1986; Nowak & Sigmund 1992; Nowak *et al.* 1995). This is realistic and it avoids the pitfall of erroneously attributing selective value to aspects of a strategy that do not come into play in an idealized, unperturbed case. For instance, in order to investigate whether reacting to low

values of a potential recipient's score is selected for, one needs to look at a situation where such scores actually occur. For this reason we will introduce errors in strategy execution in most of our analyses such that there is a probability ε for a potential donor to perform an action different from the one prescribed by the strategy.

(b) *Simulations*

We first analyse the evolutionary stability of the original image-scoring strategy 'offer help when recipient's score is at least zero', i.e. $k = 0$. Figure 1*a* shows that this strategy can be invaded by a strategy of the class that only looks to the donor's own score, namely $h = 1$. Thus, it is clear that the image-scoring strategy cannot be evolutionarily stable. Intuitively, in a population of players of the $k = 0$ strategy, an individual ought to keep his or her own score at or slightly above zero in order to exploit the aid-giving tendencies of the other group members. The evolutionary instability of the image-scoring strategy is present regardless of execution errors. In fact, for any positive cost c and at least two rounds m of interaction one readily verifies the instability of the $k = 0$ strategy.

Perhaps the most sophisticated of the strategies introduced by Nowak & Sigmund (1998*a*) is the case of $k = 0$ and $h = 1$ which belongs to the class that uses both scores. In order to give help, this strategy requires a potential recipient's score to be at least zero, but when the donor's own score is one or higher, no help is offered, regardless of the recipient's score. This strategy was found to dominate in long-term simulations by Nowak & Sigmund (see also figure 2*a*). We test the evolutionary stability of this strategy in figure 1*b*. When there are execution errors, the $k = 0$ and $h = 1$ strategy can be invaded by the $h = 1$ strategy. The intuition is as follows. In a population dominated by the $k = 0$ and $h = 1$ strategy, it is worthwhile to keep one's score at or slightly above zero. Without errors of execution there will be no negative scores in the population so that the k component of the $k = 0$ and $h = 1$ strategy does not come into play. However, as soon as there is a small frequency of negative scores, the $k = 0$ and $h = 1$ strategy suffers from sometimes failing to keep up its own score, whereas the $h = 1$ strategy always attempts to avoid negative scores.

It seems likely that the evolutionary instability of image scoring will have consequences for the kind of long-term evolution studied by Nowak & Sigmund (1998*a*). In order to clarify the issue we have performed a number of simulations. In our presentation we concentrate on the class of strategies 'offer help when own score is less than h and recipient's score is at least k' since we found this class to be the least fragile of the various proposed image-scoring strategies. Our simulations are in good agreement with previous results for a population consisting of a single social group with no errors of execution and a fairly low cost of giving help (figure 2*a*). Under the combined forces of natural selection, drift and mutation, a regime of cooperative strategies, on average dominated by $k = 0$ and $h = 1$, keeps appearing and disappearing over long stretches of time. The overall frequency of aid giving is perhaps not so high, being *ca.* 40% for the case in figure 2*a*, which would be reduced to *ca.* 30% if one were to introduce a small probability of mistake in performing actions ($\varepsilon = 0.02$). Nevertheless, the ability of the image-scoring strategies to reappear rather quickly after having been wiped out is quite interesting. As a matter of principle, the example shows that considerations of evolutionary stability need not always determine the evolutionary outcome. However, the interesting ability of cooperative image-scoring strategies to reappear depends on genetic drift. It takes tens of thousands of generations for these strategies to reappear for the island model in figure 2*b*, after which they can persist for perhaps a few thousand generations, leading to a

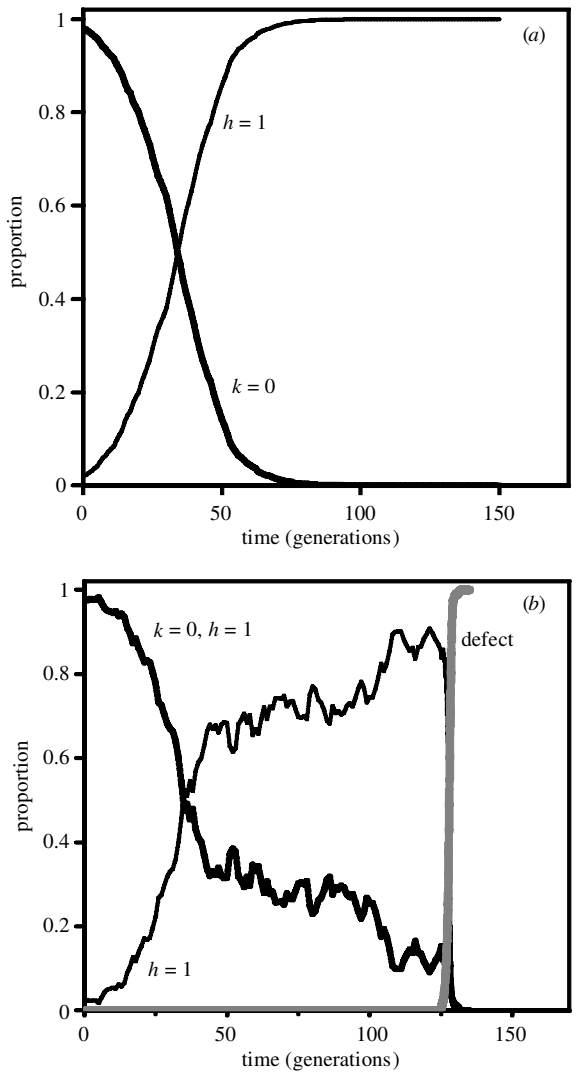


Figure 1. The evolutionary instability of image scoring. (*a*) A mutant strategy that only looks to the potential donor's own score can invade a population using an image-scoring strategy. The mutant strategy is to help when one's own score is below one ($h = 1$), regardless of the score of a potential recipient. The population strategy is to give help whenever the recipient's score is zero or above ($k = 0$), regardless of one's own score. (*b*) When there are errors in performing actions, the strategy $h = 1$ can invade a population using a strategy taking both the donor's and the recipient's scores into account ($k = 0$ and $h = 1$). Note that the $h = 1$ strategy invades but does not wipe out the $k = 0$ and $h = 1$ strategy. The resulting mixed population is vulnerable to invasion by all-out defectors. For (*a*), $\varepsilon = 0$, whereas $\varepsilon = 0.05$ for (*b*). The other parameter values are $n = 100$, $g = 100$, $m = 500$, $b = 1.0$, $c = 0.25$, $u_0 = 0$, $p = 0.9$ and $\mu = 0$.

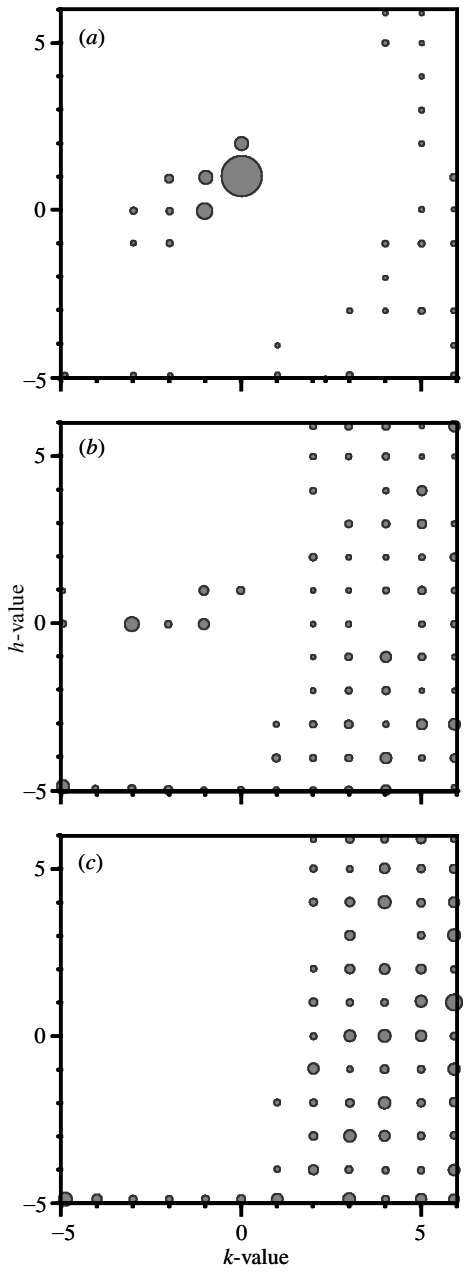


Figure 2. Long-term evolution of image scoring for three different population structures with varying influence of genetic drift. The simulated strategies are of the kind 'offer help when own score is less than h and recipient's score is at least k'. The average strategy frequencies over many generations are proportional to the area of the shaded circles. Strategies with a frequency of less than 0.5% are not shown. (*a*) For the case of a single social group ($g = 1$ and $n = 100$), as studied by Nowak & Sigmund (1998*a*), the $k = 0$ and $h = 1$ strategy tends to dominate, although strategy frequencies fluctuate over time. Averaging over 10^6 generations, help is offered in 39% of the rounds. (*b*) For an island model with limited gene flow ($g = 100$, $n = 100$ and $p = 0.9$), resulting in $F_{ST} = 0.055$, none of the strategies dominate. Averaging over 10^5 generations, help is offered in 9% of the rounds. (*c*) With stronger gene flow ($p = 0.5$), resulting in $F_{ST} = 0.015$, rather few cooperative strategies reach appreciable frequencies. Averaging over 10^5 generations, help is offered in only 2% of the rounds, which is mainly a consequence of execution errors. For (*a*), $\varepsilon = 0$, whereas $\varepsilon = 0.02$ for (*b*) and (*c*). The other parameter values are $m = 500$, $b = 1.0$, $c = 0.25$, $u_0 = 0$ and $\mu = 0.001$.

lower frequency of aid giving. With more gene flow between social groups, as in figure 2*c*, cooperative image-scoring strategies seem not to evolve at all. Thus, when the influence of genetic drift is reduced, cooperative image scoring becomes rare over the long term.

Apart from a dependence on genetic drift, we have found that the presence of cooperative image-scoring strategies over the long term depends on the cost of giving help being small in relation to the benefit of receiving help. For costs larger than in the example in figure 2, for example $c/b = 0.5$, image scoring does very badly. On the other hand, for quite small costs, for example $c/b = 0.1$, an appreciable frequency of helping through image scoring can also prevail over the long term when genetic drift is limited, which is illustrated in figure 3*a*. In addition to the low cost of helping, the initial pay-off u_0 is positive in this example, which tends to favour image scoring, perhaps because it reduces the strength of natural selection. The scope for the evolution of cooperative image scoring would then seem to be either a substantial influence of genetic drift or a very small cost of helping.

Before accepting this conclusion, let us note that the outcome of evolutionary simulations using some restricted set of strategies does not really settle the issue. If there are additional strategies that could reasonably appear in the population, adding these might change the picture. For the case of image scoring, it seems natural to also consider strategies that make use of information about how changes in one's own score would affect the probability of receiving help. For instance, consider strategies that use some estimate q_s of the current probability of receiving help as a function of a recipient's score s in the following way. A potential donor with score s evaluates the difference $q_{s+1} - q_{s-1}$ between giving help and not giving help and gives help when this difference is greater than some threshold value Δq. One possibility for the estimate is $q_s = x_s/(x_s + y_s)$, where x_s is increased by one unit for each round when a potential recipient with score s or lower receives help and y_s is increased by one unit when a potential recipient with score s or higher does not receive help. Intuitively, q_s is the proportion of rounds where a recipient with score s would have been certain to receive help among the rounds where one can deduce the outcome for recipients with score s. In order to define a starting value for q_s, we assume $x_s = 0$ for $s < 0$, $x_s = 1$ for $s \geqslant 0$, $y_s = 1$ for $s < 0$ and $y_s = 0$ for $s \geqslant 0$ at the start of a generation. This can be thought of as a prior belief that recipients with a score of zero or higher will receive help.

If we now enlarge the strategy set in figure 3*a* to also include such Δq strategies, where Δq is allowed to range between zero and one, the success of image scoring is much reduced (figure 3*b*,*c*). Some cooperative image scoring still occurs in this simulation, but it is usually wiped out fairly quickly by the spreading of Δq strategies. On average, the total proportion of Δq strategies is only 12% (figure 3*c*), but it occasionally reaches over 90% as a response to the presence of cooperative image scorers. Without cooperative image scorers, the Δq strategies are selected against since Δq players have a tendency to initially try to keep up their score.

Note that our somewhat arbitrary implementation of the Δq strategies is not likely to be maximally efficient at

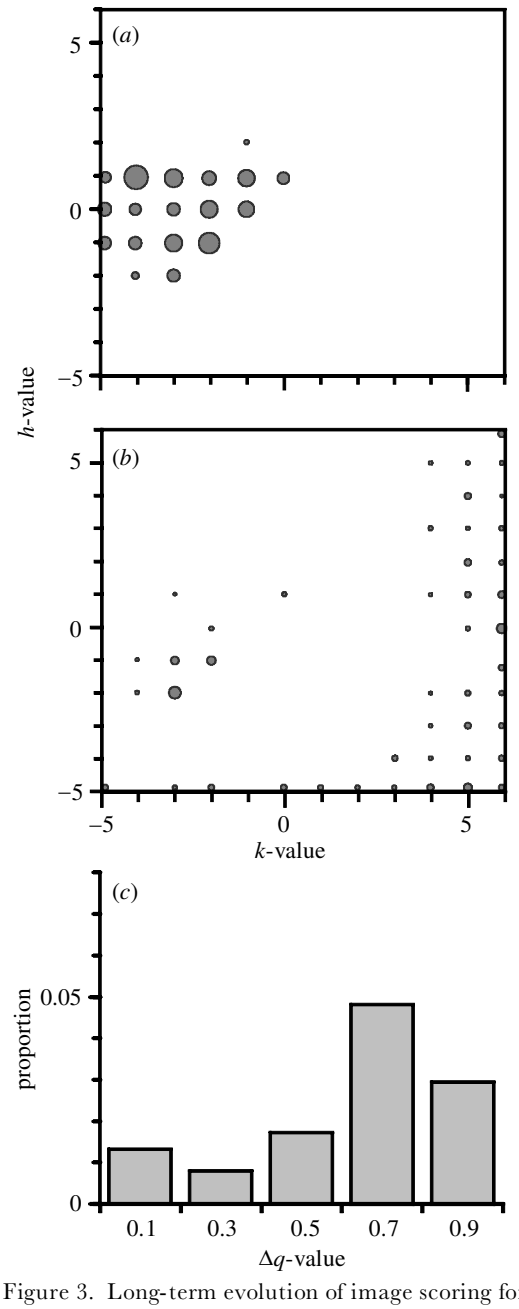


Figure 3. Long-term evolution of image scoring for two different strategy sets when the cost of helping is very small and genetic drift is limited. (*a*) For a set containing only the *k* and *h* strategies, a regime of cooperative strategies remains at fairly high frequency over the long term. The *k*- and *h*-strategy frequencies are displayed as in figure 2. Averaging over 2×10^5 generations, help is offered in 45% of the rounds. (*b*,*c*) The strategy set is extended to also include a class that can exploit image scorers. These strategies give help if it is estimated that helping would increase the probability of receiving help by at least Δq (see § 2(b)), where the threshold Δq can take the values 0.01, 0.02, . . ., 0.99. For most of the time, the level of helping stays below 5%, but occasionally cooperative image scorers reach high frequencies, only to be invaded by Δq strategies. Averaging over 2×10^5 generations, help is offered in 15% of the rounds. The distribution of the *k* and *h* strategies is shown in (*b*). These strategies make up 88% of the population and the Δq strategies account for the remaining 12%. A histogram of the distribution of the Δq strategies is shown in (*c*). The parameter values are $n = 100$, $g = 100$, $m = 500$, $b = 1.0$, $c = 0.1$, $u_0 = 5.0$, $p = 0.5$, $\mu = 0.001$ and $\varepsilon = 0.02$.

exploiting image scorers; there may well be other strategies that do better in this regard. The main point is instead that the Δq strategies exemplify psychological traits that could have been present in humans during much of their evolution. These traits include an ability to deduce how certain actions would influence the behaviour of others towards oneself and a tendency to let one's actions be shaped by such deductions. If we take into account the fact that exploiting strategies could appear even within a single generation through the application of this kind of general social intelligence rather than through the slower process of mutation and natural selection, then the scope for image scoring appears very narrow.

3. THE ROBUSTNESS OF THE STANDING STRATEGY

(a) *The model*

In studying the standing strategy, we keep our previous assumptions about the population structure, the rounds of social interaction and the inheritance of strategies. The new element is that each individual is endowed with a standing, which is good at the start of a generation. Sugden (1986) employed two variants of the concept of good standing in his analyses of reciprocity. They differ in whether an individual can regain good standing solely by helping a recipient possessing good standing or whether it is enough to help any kind of individual. Sugden used the first of these to analyse indirect reciprocity, but either works equally well. Here, we make use of the second alternative. For the structure of aid-giving interactions, Sugden considered rounds where a random group member was chosen as potential recipient and all the others were potential donors, so that several individuals could simultaneously help the recipient. There are in fact many possibilities for the assignment of potential donors and recipients, including our assumption of a single donor–recipient pair being randomly selected in a given round.

Mistakes in strategy execution are not the only kinds of errors that could realistically occur in social interactions. For instance, errors in perception of the actions used by other individuals are also likely. Errors in perception have been introduced into models of cooperation (Nowak *et al.* 1995; Boerlijst *et al.* 1997), but so far there are rather few analyses of the consequences of this type of noise. Nowak & Sigmund (1998*b*) suggested that strategies based on standing are prone to be affected by errors in perception, whereas image-scoring strategies would be less affected. In order to study the influence of errors in perception on the stability of the standing strategy, we assume there is a probability δ for an individual to misperceive an action performed by another. Clearly, these errors can lead to misperception of the standing of others, which in turn can cause an individual to be mistaken about his or her own standing. This happens when a potential donor in good standing fails to offer help to a recipient who is mistakenly perceived as lacking good standing, so that the potential donor loses good standing without being aware of it.

(b) *Evolutionary stability*

Let us investigate the stability of the standing strategy in a large population, i.e. one with many social groups,

where natural selection dominates over drift. In order to avoid the issue of kin selection, we assume global gene flow in the island model ($p = 0$). Furthermore, the assumption of a fixed number of m rounds needs to be modified to a random number of rounds in order to prevent destabilization through the end effect mentioned previously. If we let $1/m$ be the probability that the current round is the last one, the expected number of rounds becomes equal to m. With a small probability ε of execution error and with no errors of perception, a variation of Sugden's (1986) argument demonstrates when the standing strategy is a strict best reply to itself. We can distinguish two situations for an individual selected as a potential donor. First, if the individual is in good standing but the potential recipient is not, the action used will not influence future pay-offs. It would then be suboptimal to waste the amount c by offering help, so the individual should defect. Second, for all other combinations of standing, the individual must offer help to be in good standing immediately after the current round. Let r be the expected number of future rounds where the individual acts as a recipient before either becoming a donor again or the ending of the game. A rather simple computation produces $r = (m-1)/(n+m-1)$. Helping will be better than defecting if $rb - c > 0$, which is thus the condition for the standing strategy to be a strict best reply to itself. According to Maynard Smith's (1982) first criterion, the strategy is then an evolutionarily stable strategy.

Suppose now that there is also a small probability of misperception so that ε and δ are both small but positive. We will present a simplified argument that only considers first-order effects of misperceptions. In this argument, we also assume a fairly large group size n so that a given donor–recipient pair typically interacts once only. First, look at the situation where an individual chosen as a potential donor perceives him- or herself to be in good standing but perceives the potential recipient as lacking good standing. Let v be the probability that the individual's perception of the situation is mistaken. This probability is given by $v = \delta/(\varepsilon + \delta)$ for the most common case of a single perceived previous defection by the potential recipient. An individual who is mistaken but defects will lack good standing immediately after the current round and will then lose the expected amount rb from not receiving help. Since defecting saves the amount c, the individual should defect when $c - vrb > 0$. Second, for all other combinations of standing, the individual will typically either be right about the recipient being in good standing or about him- or herself not being in good standing. The individual should then offer help when $rb - c > 0$. Putting the two inequalities together we obtain the stability condition $vrb < c < rb$. For small v, i.e. if δ is considerably smaller than ε, this approaches our previous condition. On the other hand, if δ is sufficiently large compared to ε, so that v is sufficiently close to 1, or if c is small enough, unconditional cooperation can invade the standing strategy. Let us emphasize that our treatment has ignored certain rare occurrences, e.g. an individual that is chosen repeatedly as a potential donor to potential recipients perceived as lacking good standing, for which the probability v will be larger. In addition, the effects of misperceptions could be more serious for very small group sizes n.

(c) *Simulations*

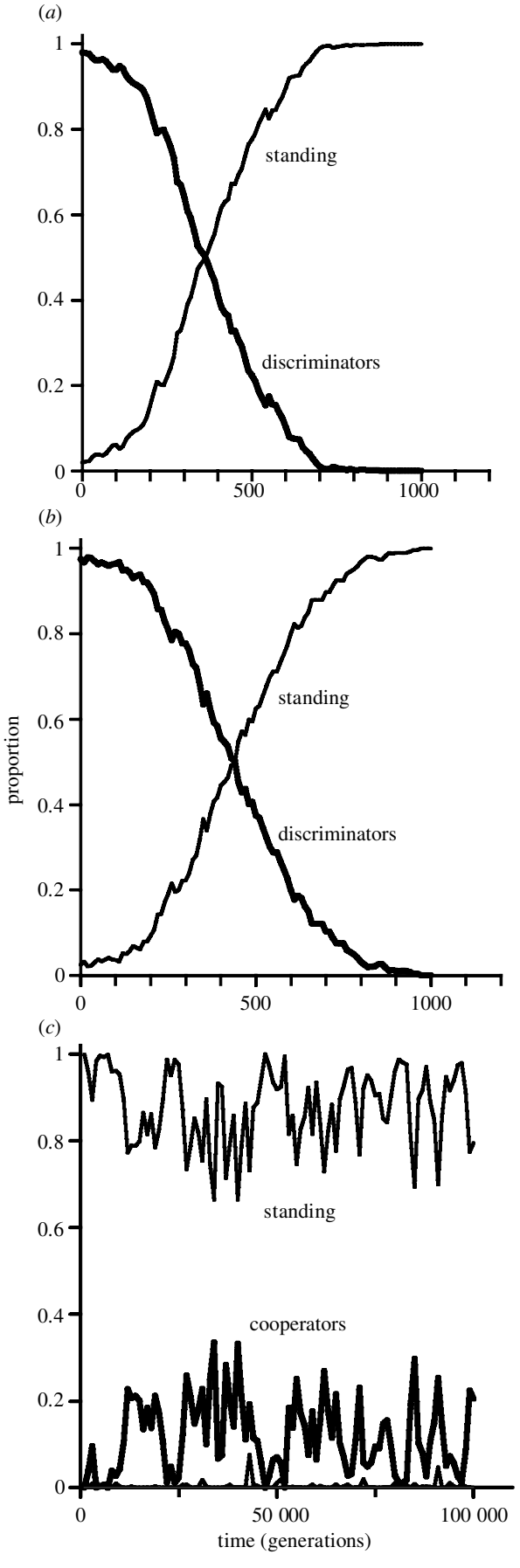
In order to see how the standing strategy performs in comparison with image scoring when there are errors in perception, we performed a number of simulations. We made use of a small set of image-scoring strategies for which the scores were restricted to zero and minus one. These were analysed in detail by Nowak & Sigmund (1998*b*) who referred to the strategies $k = -1$, $k = 0$ and $k = 1$ as cooperators, discriminators and defectors. The discriminator strategy is somewhat similar to the standing strategy, but differs by having to pay the price of receiving a negative score when refusing to help a potential recipient with a negative score.

When there are errors of execution but no errors of perception, the standing strategy can invade and take over a population of discriminators, as illustrated in figure 4*a*. Considering errors of perception also, the picture is essentially unchanged: the standing strategy takes over a population of discriminators (figure 4*b*). Our stability condition for the standing strategy is not fulfilled for the parameter values in figure 4*b*, which result in $v = 0.5$ and $r = 0.833$. Nevertheless, the standing strategy persists over the long term in a mixture with cooperators (figure 4*c*). We also found that, with a smaller influence of errors of perception (e.g. $\varepsilon = 0.04$ and $\delta = 0.01$), the standing strategy persists at a very high frequency, in agreement with our stability condition. We tried other sets of alternative strategies, such as the k and h strategies in figure 3, with similar results. Although this does not fully resolve the issue, it seems that the standing strategy also has some robustness when there are errors in perception.

(d) *State-dependent reciprocity*

So far we have assumed that all members of a social group have identical characteristics, for instance they have the same ability to give aid. This assumption is idealized and suffers from the drawback of ignoring a potentially important factor causing variability in aid-giving behaviour. If some individuals have little to offer or would find it very expensive to give help, one might expect them to refrain from helping. Other individuals may be in a better position and be more willing to offer help. The variability in behaviour caused by quality differences between individuals is analogous to the variability caused by mistakes in executing actions. However, with quality differences there is the possibility that aid-giving behaviour could serve to communicate individual quality. This kind of situation was analysed by Leimar (1997) for direct reciprocity.

The standing strategy has an interesting connection to cooperative strategies that are based on communication of individual quality. To demonstrate the connection, assume each individual has a private state z, which is known to the individual him- or herself but is not obvious to others. These states are meant to represent variation in the ability to give help. For simplicity, let there be two states, $z = 0$ and $z = 1$, with the cost of giving help being c_z in state z. If $c_1 < c_0$, the state $z = 1$, where the cost of giving help is smaller, will correspond to high quality. We allow for the possibility that the private state might change over time. Following Leimar (1997), we assume that z is stochastic with a rate α of transition from $z = 1$ to



$z = 0$ and a rate β of transition from $z = 0$ to $z = 1$. With no further information, the probability of finding an individual in the high-quality state is then $w = \beta/(\alpha + \beta)$. The rates of transition could be substantial, so that individuals often change private state or the rates could be very low.

Let each individual be endowed with standing. Just as before, an individual is in good standing initially and loses good standing when failing to help a potential recipient in good standing, but regains good standing after offering help. Consider the following strategy of state-dependent, indirect reciprocity. When in the low-quality private state, do not offer help. When in the high-quality private state, offer help as a potential donor when the standing strategy would offer help, i.e. when not in good standing or when the potential recipient is in good standing. Before going into the question of evolutionary stability, note that the strategy implies that an individual offering help reveals him- or herself as being of high quality. In addition, when the individual fails to offer help, although the standing strategy would prescribe helping, the individual reveals him- or herself as being of low quality. It is now easy to see that, for an individual in good standing, the other group members would estimate the probability of the individual being in the high-quality private state as greater than or equal to the *a priori* value w, whereas they would estimate the probability as less than w for an individual not in good standing. Thus, standing can be viewed as an assessment of individual quality.

Generally speaking, state-dependent reciprocity can be evolutionarily stable when c_1 is small enough and c_0 is large enough in relation to the benefit b of receiving help. For instance, if $c_0 > b$ it will never be worthwhile offering help when in the low-quality private state. Consider the same kind of large population as in the reasoning above for the stability of the standing strategy, but ignore errors of execution and perception. Suppose members of the population are using the strategy of state-dependent, indirect reciprocity. If the probability of being in the high-quality private state is quite high, so that w is close to 1, it will clearly be optimal to act according to the

Figure 4. The performance of the standing strategy against discriminators, unconditional cooperators and defectors. Discriminators use a $k = 0$ image-scoring strategy where the range of scores is restricted to zero and -1. (*a*) When there are errors of execution ($\varepsilon = 0.05$) the standing strategy can invade a population of discriminators. (*b*) When there are both errors of execution and errors of perception ($\varepsilon = 0.025$ and $\delta = 0.025$) the standing strategy can still invade a population of discriminators. (*c*) Long-term evolution of the standing strategy when there are both errors of execution and errors of perception ($\varepsilon = 0.025$ and $\delta = 0.025$). The alternative strategies are cooperators, discriminators and defectors. The standing strategy dominates over the long term, but cooperators are also present at appreciable frequencies. Discriminators occasionally reach around 5% (finer curve at the bottom of the graph), but defectors stay at frequencies below 1%. For these conditions, the standing strategy is thus not evolutionarily stable, but seems to have an ability to persist in a mixture with unconditional cooperators. The rate of mutation is $\mu = 0$ in (*a*) and (*b*) and $\mu = 0.0001$ in (*c*). The other parameter values are $n = 100$, $g = 100$, $m = 500$, $b = 1.0$, $c = 0.25$, $u_0 = 0.0$ and $p = 0.9$.

standing strategy when in this state, provided that $rb - c_1 > 0$. Thus, we are essentially back in the previous situation, but with rare occurrences of the low-quality private state playing the role of mistaken actions. If the low-quality private state is more common, a random group member is less likely to offer help and the cost c_1 needs to be smaller for helping to be worthwhile. For large group size, where the actual proportion of high-quality individuals will stay close to the expected proportion w, the condition becomes $wrb - c_1 > 0$, but otherwise one needs to take into account fluctuations in group composition. We refrain from going into more detail here.

When changes in individual quality are not too frequent, the actions of the strategy of state-dependent reciprocity will bring about a subgroup within the social group, consisting of high-quality individuals who preferentially assist each other. Thus, one could interpret state-dependent reciprocity as leading to coalition formation where individuals with certain abilities join together. Being in good standing would then indicate membership in such a coalition.

If acts of giving reveal important aspects of individual quality, there is the possibility that this information, which is reliably signalled in this way, could be used in other contexts. For instance, the ability to provide assistance could influence mate choice, as well as the formation of other types of partnerships. This could promote the evolution of state-dependent reciprocity since there would be additional benefits of helping, apart from being helped in return. On the other hand, if the value of being helped in return becomes very small compared to other benefits of displaying high quality, there would be less of a reason to direct the assistance preferentially towards individuals of good standing. This would lead to aid giving as a quality signal, but without reciprocity. Zahavi (1977, 1995) argued that helping in social groups functions in such a non-reciprocal manner.

4. DISCUSSION

When analysing questions of evolutionary stability, the traditional modelling style is to treat natural selection as dominating over other forces, such as genetic drift or mutation pressure, which could also cause evolutionary change. For instance, stability criteria are usually formulated solely in terms of fitness differences (Maynard Smith 1982). Another traditional element in arguments about evolutionary stability is to use the largest feasible set of alternative strategies. The main motivation for this style of modelling is not that it is always realistic to regard natural selection as dominating or that any conceivable phenotype would be produced by mutation in a reasonable span of time, but rather that it leads to a transparent and definite argument.

Computer simulation of evolutionary processes is a contrasting modelling style where natural selection, drift and mutation pressure can all influence the course of a simulation. The phenotype sets used in simulations are necessarily restricted, often to only a handful of alternatives. The main advantage of simulation and the likely reason for its increasing popularity is that the method can handle complex situations that otherwise might defy analysis. However, the advantage is bought at a considerable cost of limited understanding of the factors responsible for an evolutionary outcome. This is all the more serious when evolutionary simulations deal with highly stylized situations and aim to make conceptual points rather than incorporate all the details of real evolutionary processes. A reasonable attitude towards evolutionary simulations could then be to accept them as potentially useful, but to view results obtained in this way with a fair amount of scepticism when it is not entirely clear what factors were responsible for the results.

With regard to image scoring as a mechanism for indirect reciprocity (Nowak & Sigmund 1998*a*), one can note that the simulations aimed at analysing the most basic image-scoring strategy, i.e. to offer help when the recipient's score is at least zero ($k = 0$), were performed with an overly restricted set of strategies. The strategies used were to offer help when the recipient's score is at least k, with k varying from -5 to 6 and did not include very simple alternatives that might efficiently exploit the $k = 0$ strategy, such as the strategy of offering help when the donor's own score is less than 1 ($h = 1$) (see figure 1*a*). The analysis of image scoring by Lotem *et al.* (1999) also suffered from using this restricted strategy set. Little can be concluded from simulations where even the most natural invaders of the supposedly dominating strategies are absent.

In the analysis by Nowak & Sigmund (1998*a*) of the k- and h-strategy set (offer help when recipient's score is at least k and own score is less than h) the situation is quite different. Although the supposedly dominating strategy $k = 0$ and $h = 1$ can be invaded by $h = 1$ and the resulting mixture can be wiped out by defectors (figure 1*b*), all these strategies were in fact part of the strategy set used in the simulation. In our opinion, the results of Nowak & Sigmund (1998*a*), which were corroborated by our own simulations (figures 2*a* and 3*a*), point to a potentially quite interesting evolutionary process. However, because of its dependence on strong genetic drift (figure 2) or very low cost of giving (figure 3), we doubt that this process has any relevance for indirect reciprocity in humans. As illustrated in figure 3, there is also the issue of considering additional, more complex strategies as alternatives.

In contrast to image scoring, Sugden's (1986) standing strategy is a rather robust implementation of indirect reciprocity. The standing strategy is similar to image scoring in that help is preferentially channelled towards individuals who have themselves been generous, but has an additional characteristic in that there is no loss of good standing for someone who avoids helping potential recipients lacking good standing. At present, it is not known whether possible cases of human indirect reciprocity display this characteristic. Wedekind & Milinski (2000) showed experimentally that there is a tendency for people to aid recipients who have been generous to others in earlier interactions. Their experiment left open the question of to what extent a refusal to help an unhelpful individual is held against a potential donor in comparison to a refusal to help a helpful one.

Based on modelling considerations, there is little doubt that indirect reciprocity, in one form or another, could evolve, provided that a reasonably fair and efficient mechanism of assigning donors and recipients is already

in place. However, the latter requirement is far from easy to satisfy and may represent a major obstacle for the evolution of indirect reciprocity. Any stable implementation of indirect reciprocity would seem to presuppose a well-organized society, with a fair amount of agreement between its members as to which circumstances define the roles of donor and recipient. For instance, Sugden (1986) used the illustration of 'friendly societies' and 'sick clubs' run by working men in 19th-century England where members paid weekly subscriptions and received benefits during periods when they were too ill to work. These kinds of arrangements might well appear as a result of cultural processes, but it is harder to see the role of genetic evolution.

Zahavi (1977, 1995) suggested a very different type of explanation of seemingly altruistic behaviour in social groups. He argued that offering expensive help is an uncheatable indicator of a high-quality individual and that such acts should be seen as competitive rather than cooperative. High-quality individuals give help in order to gain prestige and the reproductive advantages that follow from a high-rank position. According to Zahavi, helping does not occur between high-quality individuals, which would be expected from our model of state-dependent reciprocity, but rather it occurs from a high-ranked individual to a low-ranked one. The logic of Zahavi's argument seems flawless, although it is perhaps not clear why individual quality should be demonstrated through altruism rather than in some other way. At the least, casual observation makes it easy to believe that competitive altruism exists as a human trait. State-dependent reciprocity, where individuals with certain skills or abilities preferentially assist each other, would also seem to be a human trait. In either case, it will be difficult to determine whether such traits are evolved adaptations rather than primarily cultural traits.

This work was supported by the Swedish Natural Science Research Council and by the Deutsche Forschungsgemeinschaft.

REFERENCES

Alexander, R. D. 1979 *Darwinism and human affairs*. Seattle, WA: University of Washington Press.

Alexander, R. D. 1987 *The biology of moral systems*. New York: Aldine de Gruyter.

Boerlijst, M. C., Nowak, M. A. & Sigmund, K. 1997 The logic of contrition. *J. Theor. Biol.* **185**, 281–293.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1996 *The history and geography of human genes*. Princeton University Press.

Kimura, M. & Weiss, G. H. 1964 The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576.

Leimar, O. 1997 Reciprocity and communication of partner quality. *Proc. R. Soc. Lond.* B **264**, 1209–1215.

Lotem, A., Fishman, M. A. & Stone, L. 1999 Evolution of cooperation between individuals. *Nature* **400**, 226–227.

Luce, R. D. & Raiffa, H. 1957 *Games and decisions*. New York: Wiley.

Maynard Smith, J. 1982 *Evolution and the theory of games*. Cambridge University Press.

Nowak, M. A. & Sigmund, K. 1992 Tit for tat in heterogeneous populations. *Nature* **355**, 250–253.

Nowak, M. A. & Sigmund, K. 1998*a* Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577.

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

Nowak, M. A., Sigmund, K. & El-Sedy, E. 1995 Automata, repeated games and noise. *J. Math. Biol.* **33**, 703–722.

Selten, R. & Hammerstein, P. 1984 Gaps in Harley's argument on evolutionarily stable learning rules and in the logic of 'tit for tat'. *Behav. Brain Sci.* **7**, 115–116.

Sugden, R. 1986 *The economics of rights, co-operation and welfare*. Oxford, UK: Basil Blackwell.

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

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

Wright, S. 1943 Isolation by distance. *Genetics* **28**, 114–138.

Zahavi, A. 1977 Reliability in communication systems and the evolution of altruism. In *Evolutionary ecology* (ed. B. Stonehouse & C. M. Perrins), pp. 253–259. London: Macmillan Press.

Zahavi, A. 1995 Altruism as a handicap—the limitations of kin selection and reciprocity. *J. Avian Biol.* **26**, 1–3.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.