*Erschienen in: Games and Economic Behavior ; 67 (2009), 2. - S. 399-407*
*https://dx.doi.org/10.1016/j.geb.2008.12.006*

# Indirect reciprocity and strategic reputation building in an experimental helping game ☆

Dirk Engelmann [a,*,1], Urs Fischbacher [b,c]

[a] *Department of Economics, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK*
[b] *Department of Economics, University of Konstanz, Box 131, 78457 Konstanz, Germany*
[c] *Thurgau Institute of Economics, Hauptstrasse 90, 8280 Kreuzlingen, Switzerland*

A B S T R A C T

We study indirect reciprocity and strategic reputation building in an experimental helping game. At any time only half of the subjects can build a reputation. This allows us to study both pure indirect reciprocity that is not contaminated by strategic reputation building and the impact of incentives for strategic reputation building on the helping rate. We find that pure indirect reciprocity exists, but also that the helping decisions are substantially affected by strategic considerations. Finally, we find that strategic do better than non-strategic players and non-reciprocal do better than reciprocal players, casting doubt on previously proposed evolutionary explanations for indirect reciprocity.

*Keywords:*
Indirect reciprocity
Reputation
Experimental economics

## 1. Introduction

Among the recent approaches to conceive a more realistic model of human behavior by extending economic theory by aspects that go beyond narrow self-interest, reciprocity has been prominent, both in theoretical (e.g. Rabin, 1993; Levine, 1998; Fehr and Schmidt, 1999; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) and experimental work (e.g. Berg et al., 1995; Fehr et al., 1993). The literature has so far focussed primarily on direct reciprocity, where a person who is affected by the choice of another person can directly reward or punish the latter. Often, though, it is not possible to reward or punish a person directly. Thus, the focus of our experimental study is indirect reciprocity, where friendly or hostile acts of one person towards another are rewarded or punished by a third party. To enable a third party to punish or reward, the information about the first person's decision has to be transmitted to the third party. Thus, indirect reciprocity is closely linked to reputation and status. This is also the view of Alexander (1987) who introduced the term indirect reciprocity.

☆ We thank Tim Cason, James Cox, Alan Durell, Charles Efferson, Ernst Fehr, Simon Gächter, Lorenz Götte, Werner Güth, Holger Herz, Steffen Huck, Michael Kosfeld, Manfred Milinski, Wieland Müller, Michael Näf, Ronald Oaxaca, Axel Ockenfels, Andreas Ortmann, Arno Riedl, Rupert Sausgruber, Daniel Schunk, Dirk Semmann, Viatcheslav Vinogradov, Ferdinand von Siemens, Georg Weizsäcker, and Christian Zehnder as well as the associate editor and two anonymous referees for helpful comments and suggestions and Richard Stock for editing. Urs Fischbacher acknowledges support from the Swiss National Science Foundation (project number 1214-05100.97), the Network on the Evolution of Preferences and Social Norms of the MacArthur Foundation, and the EU-TMR Research Network ENDEAR (FMRX-CTP98-0238). Dirk Engelmann acknowledges financial support from the *Deutsche Forschungsgemeinschaft* (DFG, grant No. EN 459/1) and from the institutional research grant AV0Z70850503 of the Economics Institute of the Academy of Sciences of the Czech Republic, v.v.i.

* Corresponding author. Fax: +44 1784 439534.
*E-mail addresses:* dirk.engelmann@rhul.ac.uk (D. Engelmann), urs.fischbacher@uni-konstanz.de (U. Fischbacher).

[1] Dirk Engelmann is also an Affiliate of the Centre for Experimental Economics at the University of Copenhagen and a Senior Researcher at the Economics Institute of the Academy of Sciences of the Czech Republic, v.v.i.

According to him, indirect reciprocity creates an incentive for friendly behavior and, thus, provides the evolutionary basis for moral systems prescribing cooperation.

For such a system of cooperation based on indirect reciprocity to work, two conditions have to be satisfied. First, people have to be rewarded for good reputation (i.e. enough others have to act in an indirectly reciprocal way) and second, they have to be willing to invest into reputation (i.e. they have to be aware that others act in an indirectly reciprocal way). As evidence for the first, Milinski et al. (2002) have shown that in an experiment donations to UNICEF are rewarded by other players. Harbaugh (1998) argues (and provides supporting field data) that donations to charity are in part driven by a prestige motive,[2] which supports the second of the above conditions. Apparently charities are aware of the prestige motive (which might be driven by expectations of indirect reciprocity) since it is common practice to announce donors' names and contributions. The interplay of indirect reciprocity and strategic reputation building can thus have substantial impact on economically relevant interaction.

Indirect reciprocity might be most important in mid-size groups with good information channels such as small towns, political parties, or business networks. In such groups, direct interaction is likely to be infrequent, so the scope for direct reciprocity is limited. On the other hand, reputation reaches potential interaction partners in sufficient quality such that a cooperative system based on indirect reciprocity can develop. In smaller groups, repeated interaction is frequent, so that cooperation can already result from direct reciprocity alone. In larger groups, information will frequently be too noisy and diluted before it reaches future interaction partners, limiting the impact of indirect reciprocity.

Seinen and Schram (2006) have conducted an experimental helping game to explore indirect reciprocity. In this game, players are randomly matched and assigned to the role of a donor and a recipient. The donor can help the recipient at a cost smaller than the recipient's benefit. A subject's previous helping decisions as donor are stored in a so-called image score and the recipient's score is presented to the donor before he decides whether to help or not. This game is nicely suited to study indirect reciprocity because it precludes (in anonymous and sufficiently large groups) any effects of direct reciprocity as opposed to games such as the prisoner's dilemma. Seinen and Schram (2006) find evidence of indirect reciprocity, because many donors base their helping decision on the image score of the recipient.[3] A substantial part of the donors, however, also base their decision on their own image score, indicating that strategic reputation building is a major force as well. The problem here is that any player whose choice can be indirectly reciprocal is at the same time influencing his own reputation. Thus, when player A helps player B who has a good reputation, we cannot be entirely sure whether this was done to reward B or to boost A's own reputation.

To assess the interplay of indirect reciprocity and strategic reputation building, it is necessary to separate out strategic incentives when observing indirectly reciprocal actions. An experimental design that achieves this aim has to allow us to identify whether observed helping choices can be influenced by strategic reputation building or not. When indirect reciprocity is not contaminated by incentives for strategic reputation building by the donor, we call this pure indirect reciprocity.[4] In Seinen and Schram the helping decision can always be driven partly by the goal to achieve a high score to receive future rewards. To disentangle indirect reciprocity and strategic reputation building, we use a helping game where in any period only half of the players have "public" scores that are seen by donors, while the other players have "private" scores. In particular, each subject has a public score either in the first 40 periods of the experiment or in the last 40 periods. This allows us to identify the effects of strategic reputation building and whether there is any pure indirect reciprocity. First, since donors with private scores interact with recipients with public scores, we can study pure indirect reciprocity.[5] Second, by comparing the behavior of donors with public and private scores, we can evaluate the relative impact of strategic reputation building on the helping rates.[6]

[2] Andreoni and Petrie (2004) provide experimental evidence on the prestige motive. They show that subjects, when having the options to contribute both to an anonymous and a broadcast public good, overwhelmingly choose the latter.

[3] Wedekind and Milinski (2000) provide the first experimental test of indirect reciprocity, based on only six periods. They find support for indirect reciprocity in the sense that recipients who are helped have had higher scores on average than recipients who are not helped. Furthermore, donors who rarely help rather do so when the recipient has a high image score.

[4] Note that we understand indirect reciprocity in the sense of indirectly reciprocal actions, that is actions that reward somebody who has been kind to a third party or punish somebody who has been unkind to a third party. One, but not the only, possibly underlying motivation might be a desire to be indirectly reciprocal, that is a player might directly derive utility from behaving in an indirectly reciprocal way. As we argue below, models of reciprocal motivation capture behavior in our experiment better than competing models. What is actually the route why players behave in an indirectly reciprocal way is, however, not our main concern. The crucial issue is that we control for strategic incentives. Thus pure indirect reciprocity means here that donors condition on recipients' reputation when they cannot influence their own reputation.

[5] People might also help when they have a private score because they believe to be observed by the experimenter or because the setting activates psychological mechanisms adapted to situations where reputation building is crucial (see e.g. Hagen and Hammerstein, 2006). This is an issue in all experiments where participants can observe others' choices, but it could be particularly important in our experiment because of the swap of roles of players with public and private score. This might make them more aware of the fact that their score can be observed—at least by the experimenter, or lets them internalize reputation building. Thus they might want to built a good reputation in the eye of the experimenter, as is suggested by the results of Hoffman et al. (1996) that dictator game giving is significantly decreased if the experimenter cannot attribute choices to individual participants. However, these explanations provide only an alternative to players being indirectly reciprocally motivated. They do not question that players act purely indirectly reciprocally, i.e., they help in the absence of actual strategic incentives.

[6] In the small town example, our players with a private score correspond to short-term visitors who spent just enough time in the town to pick up the local gossip about their respective interaction partners, but not long enough in order to have word about their own actions spread around.

We test three main hypotheses. First, indirect reciprocity is present, i.e. the probability that donors help increases in recipients' image scores. In particular, pure indirect reciprocity is present, i.e. we will find this also when subjects do not have strategic incentives to build a reputation. Second, subjects strategically build a reputation, i.e. for any given score of the recipient (including a private score) the average helping rate of donors with public image scores is higher than that of donors with private scores. Third, strategic reputation building weakens the reciprocal relation, i.e. the dependence of the donor's helping rate on the recipient's score is weaker for donors with public scores than for donors with private scores. We find support for all three hypotheses. There is a clear positive relation between helping rates and recipients' scores for both donors with public and private scores. The latter provides evidence for pure indirect reciprocity, and this is, to the best of our knowledge, the first from a laboratory experiment. The average helping rate of donors with public scores is, however, more than twice the average helping rate of donors with private scores. Hence, strategic reputation building plays an important role as well. Furthermore, strategic reputation building undermines indirect reciprocity. The probability to help increases significantly less in the recipient's score for donors with public scores than for those with private scores.

Our experiment also provides a test for models of the evolution of human cooperation. Looking for explanations for the existence of indirect reciprocity, Nowak and Sigmund (1998) have conducted simulations of an evolutionary process based on a repeated helping game. They find that maximally discriminating players will eventually take over the population. Leimar and Hammerstein (2001), however, show that this result is based on a too restricted initial set of available strategies. Subjects who are not indirectly reciprocal but only help in order to keep their own score at a level that induces a high probability of being helped (and hence base their decision only on their own score), could invade and take over a population of players who base their choice only on the recipient's score.[7] In our experiment about 15% of the population are pure strategists who are not reciprocal. Furthermore, these subjects obtain a higher material payoff, which is consistent with the invasion argument by Leimar and Hammerstein (2001) and casts some doubts on the evolutionary explanation for indirect reciprocity suggested by Nowak and Sigmund (1998).

The paper proceeds as follows. Section 2 presents the helping game and the experimental design. The results are presented in Section 3 and possible explanations are discussed in Section 4. Section 5 summarizes our results and provides concluding remarks.

## 2. Experimental design and procedures

### *2.1. The helping game*

We conducted a computerized repeated helping game similar to the game studied by Nowak and Sigmund (1998) and Seinen and Schram (2006). There were 16 subjects in each of our five experimental sessions. The helping game was repeated for 80 periods. In each period the subjects were randomly matched (independently between periods) in pairs and the role of donor and recipient were randomly assigned. The donor had the choice whether or not to help the recipient at a cost $c$ of 6 "Points," which yielded a benefit $b$ of 15 Points for the recipient. The recipient had no choice to make.

Each subject had a public score either in the first 40 periods or in the last 40 periods. All subjects were informed about this before the start of the experiment. The common knowledge of this change of roles ensured that subjects were in a symmetric position (at least over the whole course of the experiment). Hence, it precluded that donors with public and private scores behaved differently because they considered themselves advantaged or disadvantaged. Thus, we can clearly attribute behavioral differences between donors with public and private scores to strategic incentives. A score consisted of the number of times the subject had helped and had not helped in the last 5 times as a donor. In case the subject had so far been in the role of the donor less than 5 times, the score consisted of the total number of help and not help decisions so far. When the recipient had a public score, the donor was informed about this score before making the decision to help. A subject with a public score was also informed about her or his own score. In case the subject had a private score, no score information was displayed (but subjects could easily keep track of their own score and the experimental software also recorded the private scores).

A public score that is based on more than the last period allows in principle for punishments, because a player who generally helps can occasionally punish a free-rider without being punished himself if the indirectly reciprocal players do not demand a perfectly clean record. However, with our information structure, it is impossible for the subjects to distinguish punishment from occasional defection. This would require higher-order information, i.e. information about the scores of the recipients whom the current recipient did and did not help in previous periods.

The distinction into players with public and private scores is the main difference from the design of Seinen and Schram (2006), which closely implements the model of Nowak and Sigmund (1998). In their design all players had public scores, except in a control treatment without any reputation. Other differences are rather minor and consequences of the main difference. First, because of the restart with empty scores after half of the periods, there is a shorter horizon. To compensate, we reduced the scores to the last five, rather than six, decisions. Second, we chose efficiency gains in between the high and low levels used by Seinen and Schram (2006). Their treatment with high efficiency gains yields a very high helping rate,

[7] More generally, Hagen and Hammerstein (2006) argue that models of cultural evolution of cooperation often rely on conformism and can hence be vulnerable to strategic non-conformism.

which might make it difficult to detect any variation. On the other hand, we expected that the distinction into players with private and public scores would lower the helping rate. Therefore, we raised the efficiency gains above the low level in order to make an intermediate helping rate likely, which facilitates the detection of differences between players. In contrast to Seinen and Schram, we also did not choose neutral labels for the available actions (see below).

### *2.2. Experimental procedures*

The experimental software was programmed in z-Tree (Fischbacher, 2007) and the experiments were run in the computer laboratory at the Institute for Empirical Research in Economics of the University of Zurich in Fall 2001. Participants were students from a variety of fields from the University of Zurich and the Swiss Federal Institute of Technology Zurich and were recruited by phone. They were randomly assigned to cubicles in the laboratory. Written instructions were provided and participants could read through them at their own pace (see the working paper version, Engelmann and Fischbacher, 2008, for an English translation). Donor and recipient roles were labeled A and B in the instructions. The helping choices were, however, referred to as "help," because we considered the game structure so obvious, that the use of the word "help" would not invoke any interpretations that subjects would otherwise not come up with. At the end of the instructions there were five control questions to check that participants had understood the key features of the experiment. The experiment started when all participants had answered all the control questions correctly and after an oral summary of the instructions had been given.

From the second period on, subjects were informed about the outcome of the last period. The upper part of the screen reviewed their role in the preceding period, the donor's decision and the resulting payoff and total payoff so far, as well as their own score if it was public. At the same time the lower part of the screen informed the donor about his role, asked for his choice and informed him either about the public score of the recipient or that the recipient had a private score. It only informed the recipient about his role and that he did not have to make a choice. Following period 40, the roles of subjects with public and private scores were switched and the scores were cleared.

At the end of the experiment Points were converted into Swiss Francs at a rate of 1 Point = 0.1 Swiss Francs. Subjects started the experiment with an endowment of 100 Points. No additional show-up fee was paid. The sessions took between 64 and 81 minutes and earnings ranged from 6.40 to 55.60 Swiss Francs with an average of 29.36 Swiss Francs (including the initial endowment of 10 Francs).[8]

## 3. Experimental results

The overall experimental results are displayed in Fig. 1, which shows the average helping behavior of donors with public and private scores for different public recipient scores.[9] The average helping rates for the individual sessions by score status of donors and recipients are presented in Table 1 (for all scores of the recipients aggregated). Table 1 shows in particular that helping rates are quite high (32%) even when both donor and recipient have a private score, i.e. in a situation where indirect reciprocity and strategic reputation building cannot play a role. This suggests that non-selfish motives such as unconditional altruism or efficiency concerns play a role as well. We can infer from Fig. 1 and Table 1:

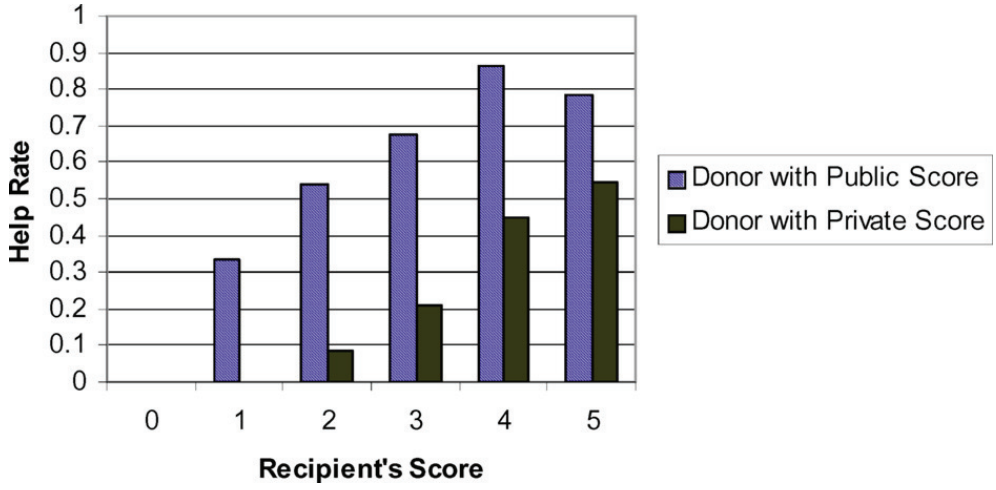


**Fig. 1.** Donors' average help rate for recipients with full public scores.

[8] At the time of the experiment, one Swiss Franc was about $ 0.61 or 0.68 Euros.

[9] We restrict the presentation of the result to recipients with full scores, i.e. to scores based on five decisions. All our results are robust to the inclusion of early periods where recipients did not yet have full scores.

**Table 1**
Average help rates by score status of donors (D) and recipients (R), recipients with full scores only.

| | Session | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| R public score | D public score | 72% | 78% | 70% | 66% | 86% | 74% |
| | D private score | 45% | 42% | 22% | 27% | 49% | 37% |
| R private score | D public score | 72% | 66% | 63% | 66% | 75% | 69% |
| | D private score | 46% | 32% | 13% | 21% | 47% | 32% |

**Result 1.** The helping rate of donors with both public and private scores increases with recipients' scores. The helping behavior of donors with private scores implies in particular that also pure indirect reciprocity, i.e., non-strategic cooperative behavior, is important.

Fig. 1 provides immediate support for the first hypothesis that donors behave indirectly reciprocally and in particular for the existence of pure indirect reciprocity. The helping rate of donors with both public and private scores clearly increases with the score of the recipient, although the relation is monotone only for donors with private scores. A straightforward statistical test confirms the significance of this positive relation. The $H_0$ hypothesis is that donors do not condition their decision on the recipient's score. We conduct a simple binomial test for this hypothesis based on the five sessions as independent observations. To obtain an estimate of whether there is a positive relation between the recipient's score and the helping probability we calculate the Spearman rank correlation between the recipient's score and the dummy variable for the donor's helping decision. Under $H_0$ in each individual session the probability that the estimated rank correlation is positive equals $\frac{1}{2}$ (actually slightly smaller, because of a small positive probability for zero correlation). We find a positive correlation in all five sessions and can thus reject the $H_0$ that there is no positive relation at the 5% level.[10] This holds independently of whether we consider donors with public or private scores.[11] That it holds for donors with private scores is evidence for pure indirect reciprocity.

Having a higher public score is more costly but as we have just seen also increases the likelihood of being helped. This raises the question whether there is a payoff-maximizing public score (the payoff maximizing private score is obviously 0). Table 2 shows for all sessions the net profit that results from holding specific public scores, given the observed helping rates for these public scores. It shows that in all sessions the score of 4 was the most profitable, and on average payoff increases monotonously for scores below 4.

Considering again Fig. 1 and Table 1, we make the further crucial observation:

**Result 2.** Donors with public scores help substantially more often than donors with private scores. Hence, strategic cooperative behavior is of crucial importance as well.

This provides clear support for the second hypothesis that donors strategically build a reputation. The average helping rate of donors with public scores is higher than of those with private scores for any score of the recipient (including a private score, as can be seen from Table 1).[12] The same holds for each individual session. There is only one tie, in session 1 for a score of 0. Under the $H_0$ hypothesis that strategic reputation building is not relevant, in each session the probability is less than $\frac{1}{2}$ that the helping rate is (weakly) higher for donors with public scores than for donors with private scores for any recipient score. Thus, the fact that this holds in all five sessions allows us to reject the $H_0$ at the 5% level.

Support for our first two hypotheses can also be derived from panel data analysis (with the sessions as independent units of observations). We use a random-effects probit model.[13] The first model is

$$\Pr(Help)_{it} = \Phi\big(const + \alpha \cdot RsScore_{it} + \beta \cdot DPublic_{it} + \gamma \cdot (DPublic \times RsScore)_{it}\big),$$

with $\Pr(Help)$ the probability that the donor helps, $\Phi$ the normal cumulative distribution function, *RsScore* the recipient's score, *DPublic* a dummy that is 1 if the donor has a public score and 0 otherwise, and $DPublic \times RsScore$ the interaction effect. $\alpha$ and $\beta$ are significantly positive, supporting the first and second hypothesis, $\gamma$ is significantly negative, supporting the third hypothesis (see (I) in Table 3).[14]

[10] Since the sessions are independent, the probability for an estimated positive correlation in all five sessions is (slightly smaller than) $(\frac{1}{2})^5 = \frac{1}{32} < 5\%$. The same logic will apply to all our non-parametric tests below. Since all our hypotheses are directed, we can apply one-sided tests throughout.

[11] Although for this test, we need only the sign of the correlations, we note that in all sessions the correlation was indeed significantly positive. Note, though, that the test of significance of the correlations is not a valid test because the observations are not independent.

[12] It is strictly higher except for recipients with a score of 0. There are, however, only 13 interactions with a recipient with a full score of 0. Among these, 12 are with the same subject and hence all in session 1. Furthermore, since the helping rate for donors with private scores is already 0 for recipients with a score of 1, this tie appears to simply result from censoring.

[13] All reported results are qualitatively the same for a logit model.

[14] If we restrict the analysis to donors with public score, *RsScore* is still highly significant and positive. Thus, although as predicted by the third hypothesis, the reciprocity of donors with a public score is reduced, it is not eliminated.

**Table 2**
Average expected return per period (in Points) from keeping a certain public score, based on average help rates over the whole phase with full scores.

| Score \ Session | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | −0.6 | 2.9 | 0.57 | −0.6 | −0.6 | 0.57 |
| 2 | 0.55 | 0.91 | 1.11 | 1.03 | 2.3 | 1.03 |
| 3 | 2.1 | 1.2 | 1.22 | 0.87 | 1.8 | 1.39 |
| 4 | 2.52 | 3.01 | 1.51 | 2.63 | 2.85 | 2.42 |
| 5 | 2.07 | 2.48 | 1.11 | 0.81 | 2.62 | 1.92 |

**Table 3**
Random-effects probit model for the help choice.

| | (I) | (II) |
|---|---|---|
| Recipient's score (*RsScore*) | 0.4703*** | 0.6794*** |
| | (0.0579) | (0.0761) |
| Dummy for donor with public score (*DPublic*) | 1.7319*** | 2.5351*** |
| | (0.3187) | (0.4511) |
| Interaction effect (*DPublic* × *RsScore*) | −0.1644* | −0.3569*** |
| | (0.0810) | (0.0957) |
| Donor's score (*DsScore*) | | 0.6139*** |
| | | (0.0460) |
| Interaction effect (*DPublic* × *DsScore*) | | −0.3349*** |
| | | (0.0708) |
| *const* | −2.1854*** | −4.0605*** |
| | (0.2453) | (0.3444) |
| *N* | 1135 | 1135 |
| log likelihood | −636.19 | −496.16 |

(II) includes controls for the donor's score as well as an interaction effect with the dummy whether the donor's score is public. (II) is the superior model both according to the Bayesian Information Criterion and Akaike Information Criterion. Data is restricted to the cases with full score for both players.

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

In a second regression, we also control for the donor's score and an interaction term of the donor's score and the dummy for the donor's score being public (see (II) in Table 3). This even strengthens the above results, as the absolute values of all relevant parameter estimates increase. Furthermore, the donor's score has a significant positive impact, which suggests individual differences in the propensity to help, because this implies that some donors have consistently a higher score and help more often. The interaction effect between the donor's score and the dummy indicating whether he has a public score, however, is significantly negative. This suggests that having a public score increases the helping rate more if the donor has a low score. This is consistent with strategic reputation building. This behavior is rational because the payoff maximizing score is at a high, but not the maximum possible, level.

As can be seen in Table 1, in each session for both recipients with public and private scores the helping rates of donors with public scores is about twice the helping rate of donors with private scores. Hence, the impact of strategic reputation building is not only statistically significant, but also of substantial magnitude. On the other hand, both for donors with public and private scores, the average helping rate is only slightly (about 5 percentage points) lower if the recipient has a private rather than a public score. Hence, recipients with private and public scores are on average treated nearly equally.

The importance of strategic reputation building is also very vividly illustrated by Fig. 2 which shows the distribution (absolute frequencies on top of bars) of donors' full (public or private) scores. The mode of private scores is 0, for public scores, in contrast, it is 4. Interestingly, in all sessions the score that maximizes expected payoffs for the observed helping rates is 4 (see Table 2). A probit regression supports the view that donors with public scores strategically maintain the optimal score of 4. If their score falls below 4, they increase their probability to help, if it is greater than 4, they decrease it. This is only true for donors with public scores. The helping rate of donors with private scores does not depend on whether their score is larger or smaller than 4 (see Engelmann and Fischbacher, 2008, for details).

An advantage of our design is that we can study strategic reputation building on an individual basis by comparing the helping rates with public and with private scores within subjects. We observe:

**Result 3.** There is substantial heterogeneity in behavior both in terms of indirect reciprocity and strategic reputation building.

We classify the subjects according to two dimensions. First, we call a subject strategic if her helping rate is generally higher in the part of the experiment where she has a public score than in the part where she has a private score. We find that only 20% are not strategic and slightly more than half of our subjects (43/80) are what we call strongly strategic, that is their helping rate at least doubles when they have a public score compared to a private score. Among them, 20 subjects are what we call pure strategists, who never help when they have a private score, but do so several times when they have

a public score. On a second dimension, nearly half of the subjects (39/80) are classified as reciprocal, meaning there is a clear positive relation between the recipient's score and the helping rate. Among the 43 strongly strategic players, 20 are classified as reciprocal (see Engelmann and Fischbacher, 2008, for a more detailed discussion). Interestingly, even 40% of the pure strategists are also clearly reciprocal. Hence, while their primary motive to help appears to be strategic reputation building, they are also concerned with providing incentives for the other subjects. Instead of just exploiting the cooperative system based on indirect reciprocity, they also stabilize it.

The remaining 60% of pure strategists (15% of the total population), however, appear to be of the type predicted by Leimar and Hammerstein (2001) to invade the population. In line with their prediction, we find that the strongly strategic subjects receive significantly higher payoffs than the, at most, weakly strategic and the non-reciprocal receive (albeit insignificantly) higher payoffs than the reciprocal subjects. The differences are remarkably large. In particular, the strongly strategic non-reciprocal subjects obtain on average 1.23 times the session average payoff, while the weakly or non-strategic reciprocal subjects obtain only 0.69 times the session average. This indicates that in an evolutionary game based on this repeated helping game and with the experimentally observed player types, the strongly strategic non-reciprocal types would drive out the other types and would eventually undermine cooperation. Given the large payoff differences, the evolutionary process would be quite fast for any sufficiently payoff-sensitive dynamic.

## 4. Discussion

How can we account for our results[15]? Common knowledge of selfishness and rationality implies that no help will occur in the last period for any history of play, and by backward induction, no help will occur in any period. Thus to explain our data, we need to relax the selfishness assumption or the rationality assumption.

We first note that it is commonly observed that experimental subjects do not perform many stages of backward induction and thus treat finitely but frequently repeated games like infinitely repeated games, at least in the early stages. In Engelmann and Fischbacher (2008) we show for a simplified model of an infinitely repeated helping game that there are equilibria where even selfish donors with public scores condition their choice on the recipient's score. However, any two adjacent scores that both occur with positive probability have to yield the same expected payoff, which is contradicted by Fig. 2 and Table 2. In each session, public scores of 3, 4, and 5 all occur frequently, but yield substantially different payoffs.

Thus we need to move beyond selfish preferences. Unconditional altruism or concerns for maximizing total payoff (Charness and Rabin, 2002) can explain that there is helping per se, but not that helping conditions on the recipient's score. Similarly, the model of inequality aversion by Bolton and Ockenfels (2000) does not predict any conditioning on the recipient's score, as the donor would only care about how his own payoff relates to the total payoff, but not about individual comparisons.[16] The related model by Fehr and Schmidt (1999) is based on individual comparisons and could thus capture indirect reciprocity if a recipient is more likely to be helped if he is poorer. Table 2 shows, however, that the average payoff

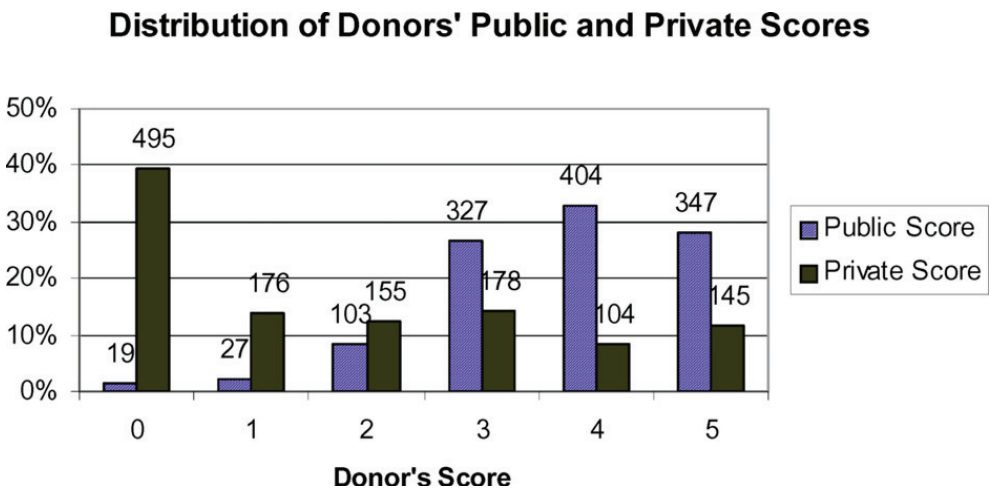


**Fig. 2.** Distribution of public and private (post-decision) donors' scores for all interactions where the donor had a full score (except for following donors' decisions in the last period because in that case the resulting score could not possibly be relevant for future interaction). Absolute numbers appear on the top of the bars. The total number of the included scores is 2480, 1227 where the score is public and 1253 where the score is private (the difference is a result of the random allocation of donor and recipient roles, apparently it just happened that players with private scores were chosen slightly more often as donors).

[15] For a much more detailed discussion of the implications of various models, see Engelmann and Fischbacher (2008).

[16] The recipient's score could be interpreted as a signal for the overall helping rate and hence the average payoff, which would lead to a higher predicted helping rate after observing a low score, the opposite of what we observe.

increases monotonously for scores below 4, while Fig. 1 shows that the helping rate increases weakly monotonously in the recipient's score, i.e. it is the players with the higher payoff that receive more help.[17]

A better explanation for indirect reciprocity is provided by reciprocity models. In the models of Rabin (1993) and Dufwenberg and Kirchsteiger (2004) player A rewards B's kindness with kindness where B's kindness is assessed by the effect B's actions have on A's payoff (specifically, B's action is kind if it leads to a payoff above A's "fair payoff"). In our setting, B's score may be partly driven by previous interactions with A, and hence serves as a noisy signal of B's kindness towards A. To increase the helping probability with the recipient's score would then be consistent with the models. In this sense, indirect reciprocity would just be a side-effect of direct reciprocity. It should then disappear in infinitely large populations. According to the models it should not weaken with a larger finite population because a recipient with the highest score still has the maximum expected kindness even if it is very unlikely, but not impossible, that he has ever helped the donor, but that may seem psychologically not very plausible.

The model by Levine (1998) appears to be more convincing in this respect, because it can capture indirect reciprocity directly. His model includes altruism and spite. Crucial for the present purpose, player A's altruism towards B is increasing in player A's estimate of player B's altruism. All else equal, a more altruistic player is more likely to help and hence B's score is positively correlated with B's altruism. Thus, indirect reciprocity, in particular the pure indirect reciprocity that we observe, is consistent with the model. In contrast to the other reciprocity models, it does not predict that indirect reciprocity will vanish in an infinitely large population. Levine's model is also consistent with a number of further observations. First, since it allows for unconditional altruism, it can capture the relatively high rate of helping by donors with private scores towards recipients with private score. Second, it is consistent with the helping rate towards recipients with private scores being similar to the average helping rate towards recipients with public scores. The population of participants with public and private scores is the same. (They are not only drawn from the same distribution, but thanks to the role reversal and seen across the whole experiment, they really are the same.) Hence the distribution of the altruism parameters is the same. Therefore, the donor's estimate of the recipient's altruism should correspond to the average altruism in the reference group (even though players with private scores actually help less often). Since the donor's choice depends only on his own altruism and his estimate of the recipient's altruism, he should thus treat a recipient with a private score like the average recipient with a public score.[18]

We note that the observed strategic reputation building is consistent with any of the models, because they all include self-interest. Since a high public score pays off, independent of the original motivation for helping, having a public score provides additional incentives to help.

We do not claim that a reciprocal motivation is the sole ultimate source for pure indirect reciprocity. There can also be strategic helping by donors with private score (in order to encourage the recipient to help in the future). However, it is not likely that this is a major force. If such strategic helping were important, it should decrease over time, but indeed helping rates do decline only in the last two periods and then primarily for the donors with public scores.[19] This late decline is quite surprising. Interestingly, the decline in helping rates is almost entirely driven by those among the subjects with public scores whom we classified as strongly strategic. This suggests the other subjects, and in particular those with private scores, do not help for strategic reasons (see Engelmann and Fischbacher, 2008, for details).

## 5. Conclusions

We have conducted an experimental helping game where at any time only half of the subjects have a public score and hence a strategic incentive to help. Thus, we can study both pure indirect reciprocity and the impact of strategic incentives. The interaction of donors with public and private scores is the fundamental difference from the helping experiment by Seinen and Schram (2006). In their experiment, all subjects could build up an image score and hence it is not possible to clearly distinguish between helping choices that are purely indirectly reciprocal and helping choices that are driven by attempts to improve one's own score.

From a general perspective, our separation between subjects who can strategically build a reputation and those who cannot provides a clean separation between strategic cooperative behavior (the difference in the behavior of donors with private and public scores) and non-strategic cooperative behavior (helping by donors with private scores). That the average helping rate of donors with private scores is more than 30% is as clear evidence for the existence of non-strategic cooperative behavior as the substantially higher average helping rate of donors with public scores is evidence for strategic reputation

[17] Naive inequality aversion, where donors ignore that other donors are more likely to help players with high score could capture the observed pattern. As our discussion of strategic behavior reveals, however, our subjects are very aware of the fact that others condition their help decision on the recipient's score.

[18] While this observation does not necessarily imply that the average help rate is exactly the same for both types of recipients, in light of this implication of Levine's model, it is not surprising that average help rates are similar.

[19] One could also argue that the positive relation between the recipient's score and the helping probability does not result from indirect reciprocity, but rather from a learning process. Donors might want to find out what is a successful score and may use the observed scores as orientation. Trying to adapt one's own score to the observed recipients' scores would imply to help when one observes a high score and not to help when one observes a low score (though this should strictly be so only in early periods or if subjects are highly myopic, because otherwise the total information one has gathered so far should dominate this period's recipient's score). This potential interpretation, however, appears to be valid only for donors with public scores, because donors with private scores do not have an incentive to find out what constitutes a successful score.

building. From a more specific perspective, we are the first to find clear evidence for pure indirect reciprocity, but we also find very strong effects of strategic reputation building. Specifically, 80% of subjects react to strategic incentives, including 25% who only help when they have an incentive to do so. The pure indirect reciprocity that we find is inconsistent with outcome-oriented models such as Fehr and Schmidt (1999) or Bolton an Ockenfels (2000). It is, in contrast, consistent with the reciprocity approaches by Rabin (1993), Dufwenberg and Kirchsteiger (2004) and in particular with the model by Levine (1998).

Our data also sheds some light on a recent discussion on the evolution of cooperation. Concerning the empirical relevance of the invasion predicted by Leimar and Hammerstein (2001), we clearly find strategic non-reciprocal players who also receive higher payoffs than other types. This casts some doubts on the evolutionary explanation for cooperation based on indirect reciprocity suggested by Nowak and Sigmund (1998) because the types predicted to undermine cooperation by exploiting the system are clearly present and more successful. Put differently, the argument that the set of potential types in the simulations is too restricted, put forward by Leimar and Hammerstein, is valid not only on theoretical grounds, but is also strongly supported by our experimental data. The exploiting types actually exist, so any simulation or evolutionary model that tries to explain altruistic behavior has to take them into account. Therefore, an evolutionary explanation for the presence of indirect reciprocity (that is documented by several experiments, including ours) has to be richer in structure to explain why reciprocal players might survive in the presence of non-reciprocal strategic players. An example would be an environment where higher-order information is available. Then the standing strategy (Sugden, 1986) can be used, which discriminates between those who punish free-riders and those who free-ride themselves. A shortcoming of the model by Nowak and Sigmund (1998) is that it does not make use of such higher-order information. Not using higher-order information in turn eases invasion.

As a final contribution, our experiment shows that evolutionary models can be tested in the laboratory, in our case by proving the existence of a type that would undermine the process that drives the result of the evolutionary model. Evolutionary explanations for a behavior are often vulnerable to the existence of strategic types that successfully mimic a property that is the basis for the evolutionary advantage of the fittest type. Exposing subjects in the laboratory to a situation as assumed by the evolutionary model permits a test for the existence of these mimicking types.

## References


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

Andreoni, James, Petrie, Ragan, 2004. Public goods experiments without confidentiality: A glimpse into fund-raising. J. Public Econ. 88, 1605–1623.

Berg, Joyce, Dickhaut, John, McCabe, Kevin, 1995. Trust, reciprocity and social history. Games Econ. Behav. 10, 122–142.

Bolton, Gary E., Ockenfels, Axel, 2000. ERC – A theory of equity, reciprocity, and competition. Amer. Econ. Rev. 90, 166–193.

Charness, Gary, Rabin, Matthew, 2002. Understanding social preferences with simple tests. Quart. J. Econ. 117, 817–869.

Dufwenberg, Martin, Kirchsteiger, Georg, 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.

Engelmann, Dirk, Fischbacher, Urs, 2008. Indirect reciprocity and strategic reputation building in an experimental helping game. Thurgau Institute of Economics Research Paper Series No. 34, http://www.twi-kreuzlingen.ch/fileadmin/pdfs/research/TWI-RPS-034-Engelmann-Fischbacher-2008-08.pdf.

Falk, Armin, Fischbacher, Urs, 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–315.

Fehr, Ernst, Schmidt, Klaus M., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 817–868.

Fehr, Ernst, Kirchsteiger, Georg, Riedl, Arno, 1993. Does fairness prevent market clearing? An experimental investigation. Quart. J. Econ. 108, 437–460.

Fischbacher, Urs, 2007. Z-tree: Zurich toolbox for ready-made economic experiments. Exper. Econ. 10, 171–178.

Hagen, Edward H., Hammerstein, Peter, 2006. Game theory and human evolution: A critique of some recent interpretations of experimental games. Theoretical Population Biology 69, 339–348.

Harbaugh, William T., 1998. The prestige motive for making charitable transfers. Amer. Econ. Rev. 88, 277–282.

Hoffman, Elizabeth, McCabe, Kevin, Smith, Vernon L., 1996. Social distance and other-regarding behavior in dictator games. Amer. Econ. Rev. 86, 653–660.

Leimar, Olof, Hammerstein, Peter, 2001. Evolution of cooperation through indirect reciprocity. Proceed. Roy. Society London: Biological Sci. 268, 745–753.

Levine, David K., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dynam. 1, 593–622.

Milinski, Manfred, Semmann, Dirk, Krambeck, Hans-Jürgen, 2002. Donors to charity gain both indirect reciprocity and political reputation. Proceed. Roy. Society London: Biological Sci. 269, 881–883.

Nowak, Martin A., Sigmund, Karl, 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577.

Rabin, Matthew, 1993. Incorporating fairness into game theory and economics. Amer. Econ. Rev. 83, 1281–1302.

Seinen, Ingrid, Schram, Arthur, 2006. Social status and group norms: Indirect reciprocity in a helping experiment. Europ. Econ. Rev. 50, 581–602.

Sugden, Robert, 1986. The Economics of Rights, Co-operation and Welfare. Basil Blackwell, Oxford.

Wedekind, Claus, Milinski, Manfred, 2000. Cooperation through image scoring in humans. Sci. 288, 850–852.