No. 34 august 2008

Dirk Engelmann
Urs Fischbacher

Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game

Research Paper Series
Thurgau Institute of Economics and Department of Economics
at the University of Konstanz

THURGAU INSTITUTE
OF ECONOMICS
at the University of Konstanz

Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game*

Dirk Engelmann[†]and Urs Fischbacher[‡]

August 19, 2008

Abstract

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. We find that the behavioral pattern can best be captured by non-selfish preferences as assumed by reciprocity models. 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.

JEL Classification: C92

Keywords: indirect reciprocity, reputation, experimental economics

*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. 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).

[†]Department of Economics, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom, dirk.engelmann@rhul.ac.uk.

[‡]Department of Economics, University of Konstanz, Box 131, 78457 Konstanz, Germany, email: urs.fischbacher@uni-konstanz.de and Thurgau Institute of Economics, Hauptstrasse 90, 8280 Kreuzlingen, Switzerland.

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, Dickhaut, and McCabe, 1995, Fehr, Kirchsteiger, and Riedl, 1993). The focus of the literature has so far been almost exclusively 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. In particular in large societies repaying a favor directly can be difficult. 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. 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, Semmann, and Krambeck (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,[1] 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

[1] 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.

reputation building can thus have substantial impact on economically relevant interaction.[2]

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

[2] Beyond indirect reciprocity, reputation building can be crucial for the functioning of markets with repeated one-shot interactions. This is increasingly relevant in markets that are becoming larger and more anonymous and hence less prone to be influenced by direct reciprocity as is exemplified by e-commerce. Bolton et al. (2004) and Keser (2002) provide experimental evidence on the importance of reputation mechanisms in environments with repeated one-shot interactions. Indirect reciprocity might be most effective in mid-size groups such as small towns, where repeated direct interaction is infrequent and hence the scope for direct reciprocity is limited but the group is small enough for reputation to spread around quickly.

[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

disentangle indirect reciprocity and strategic reputation building, we use a helping game where in any period only half of the players have a "public" score that is seen by donors, while the other players have a "private" score. 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 a private score interact with recipients with a public score, 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]

We test three main hypotheses. First, indirect reciprocity is present, i.e. the probability that donors help increases in the recipient's image score. 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 a public image score is higher than that of donors with a private score. Third, strategic

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] Helping in our design might, however, also be driven by an internalized scoring rule since subjects with a private score knew that there are subjects with a public score and hence knew that in principle somebody (for example the experimenter) could compute also their score. As a result, the pure indirect reciprocity we find could just be a result of internalizing this hypothetical score. This problem, however, will exist in any experiment, and more importantly, for any real life situation that gives an opportunity for indirect reciprocity (Hagen and Hammerstein, 2006, argue that this is a general problem in the interpretation of economic experiments designed to be anonymous). Whenever I observe somebody's previous actions before deciding how to treat him, I am confronted with the theoretical possibility that somebody could observe my action and I could hence internalize my reputation. This effect might be somewhat strengthened in our experiment because of the swap of roles of players with public and private score, but it cannot be completely excluded anyway. Moreover, this explanation provides only an alternative to players being indirectly reciprocally motivated. It does not question that they act purely indirectely reciprocally.

[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.

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 a public score than for donors with a private score. 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 a public score is, however, more than twice the average helping rate of donors with a private score. 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 a public score than for those with a private score.

Our experiment also allows us to test some qualitative hypotheses derived from various models of social preferences and thus to provide some assessment which of these models better captures crucial features of the experimental data. While inequality aversion as modeled in Fehr and Schmidt (1999) or Bolton and Ockenfels (2000) as well as efficiency seeking and maximin preferences as modeled by Charness and Rabin (2002) cannot account for important aspects of the empirical evidence, the reciprocity model by Levine (1998) is consistent with pure indirect reciprocity as well as the reputation effect. To a lesser degree this is also true for the models by Rabin (1993) and Dufwenberg and Kirchsteiger (2004), where indirect reciprocity results as a form of stochastic direct reciprocity. The reciprocity models capture the data better than the models based on distributional concerns because they consider the helping act per se as kind and thus higher scores deserve more help. The distributional models often capture reciprocity in other settings because the total payoff serves as a proxy for kindness (i.e. a low payoffs signals kindness). In this experiment, it is not a good proxy because those with a high score are not those materially worst off.

Finally, 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 image scorers (i.e. 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. In Section 3, we derive theoretical predictions for the indirect reciprocity game. The results are presented 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

[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.

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 score 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 to the design of Seinen and Schram (2006), which closely implements the model of Nowak and Sigmund (1998). In their design all players had a public score, 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 on an intermediate level of those in Seinen and Schram (2006). Their treatment with high efficiency gain yields a very high helping rate, 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, so we raised the efficiency gains beyond their low level in order to make an intermediate helping rate likely, which facilitates the detection of differences between players. In contrast to them, we also did not neutrally label 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 Appendix 2 for an English translation). Donor and recipient roles were labeled A and B in the instructions, but the helping choices were labeled as such, 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. At the same time they were either asked to make a decision or were informed that they were a recipient. 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 they had a public score in that half of the experiment. A donor was asked for his choice in the lower part of the screen and there he was either informed about the public score of the recipient or that the recipient had a private score. A recipient was only informed about his role and that he did not have to make a choice. Following period 40, the roles of subjects with public and private score 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 Franc. 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 10 Francs initial endowment).[8]

3 Predictions

In this section we consider predictions of various applicable models of behavior for our experimental setting. We first analyze the case in which all players are rational and selfish and this is common knowledge.

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

Experiments frequently find deviations from the rational selfish prediction. Our next steps address possible reasons for such deviations. The first deals with the possibility that players perceive the game as if it has an infinite horizon, which can enable indirect reciprocity and reputation building even among selfish players. Finally, we investigate the predictions of models of non-selfish preferences.

3.1 Standard Prediction

If players are selfish and rational, they will not help in the last period, irrespective of the score of the recipient. If selfishness and rationality is common knowledge, backward induction generalizes this argument to all periods. It implies that no player will ever help.

3.2 The Infinitely Repeated Game

In this subsection, we present the theoretical prediction for the infinitely repeated game. We do so because experimental evidence overwhelmingly suggests that players are not able to perform backward induction over more than a few periods and thus they might perceive the finitely (but many times) repeated game as an infinitely repeated game, at least until close to the end. We relegate a formal derivation of the infinitely repeated game to the appendix, because it is not the main focus of this paper. Here, we present the main results in an intuitive way.

In Appendix 1, we show first that there are equilibria in which players help. These equilibria exist even if players never meet a direct or indirect interaction partner again. In the prototypical equilibrium, the probability of getting help linerarly increases in one's own score and the slope of the linear function is determined by the parameters of the game - and the slope exactly makes players indifferent between helping and not helping. There is an equilibrium in which helping is independent of the donor's own score and depends only on the score of the recipient. Unfortunately, this is not true for all equilibria. There are many equilibria and there are some equilibria in which players keep their score within some interval of scores. In such an equilibrium players always help if they otherwise would get a score below this interval and they would never help if otherwise their score increases above the interval. Nevertheless, we can present some properties of any symmetric stationary equilibrium. In particular, we show that if neighboring scores occur with positive probability, players are indifferent between having one or the other score, i.e. the expected values of the scores are equal. If this were not the case, a (payoff-maximizing) player would base his help decision never on the recipient's score, but on obtaining the score that yields the higher expected payoff. In short, a selfish player conditions the helping choice on the recipient's score only if the effect on his own score does not change his expected payoff.

3.3 Social Preferences

In this and the next subsection, we discuss implications of various models of social preferences. We first analyze the incentives for the players with private score, in order to be able to disregard reputation effects. In the model by Fehr and Schmidt (1999) (FS), players dislike inequality, i.e. they dislike if other players have a higher or a lower payoff than the player herself. Applied to our experiment, a donor will only help if he has a higher payoff than the recipient, if he believes to have at least the third highest payoff, and if the donor's β (this parameter measures how much a player dislikes advantageous inequality) is unrealistically large.[9] If we assume that players perceive the game bilaterally and ignore the other players when they decide whether to help or not, the condition for helping is $\beta \geq \frac{c}{b+c} = \frac{2}{7}$, which is possible and Fehr and Schmidt derive a parameter distribution based on ultimatum game results in which about 40% of the subjects satisfy this condition. Since donors with FS preferences only help if they expect the recipient to have a lower payoff (at the end), they would condition their help on the score of the other players if the score is a signal for the payoff. This means that donors with a private score should reward recipients with scores that yield a low payoff. Furthermore, players with higher payoffs are more likely to help.[10]

In the model of Bolton and Ockenfels (2000) (BO), players dislike inequality in comparison to the group average. Therefore, donors could help if they believe to earn more than the average of all other players.

[9] With α and β the FS parameters of the donor's aversion to negative and positive inequality, let r be the number of players who are richer than the donor. For simplicity assume further that the donor's decision does not change the rankings of the players. Then the donor helps if $\beta \geq \frac{c(N-1+r\alpha)}{c(N-1-r)+b}$. Because $\beta < 1$, this condistion can only be satisfied if $r < 3$. Furthermore, the lowest possible value of β equals $\frac{6}{7}$.

[10] Subjects are predicted to help more often recipients with higher scores only if they are naive, i.e. if they evaluate equity only based on the cost of helping and ignore the benefit of receiving help in the future.

However, there is no reason for BO players to favor particular other players, e.g., based on the score. The recipient's score could be interpreted as a signal for the overall helping rate and hence the average payoff, which would imply a higher predicted helping rate after observing a low score, the opposite of an indirectly reciprocal helping behavior.

Charness and Rabin (2002) suggest a model in which players care about efficiency and about the payoff of the poorest player. If players' concern for efficiency is sufficiently high, they should always help. The concern for the income of the poorest yields similar predictions as the FS model.

In the reciprocity models of Rabin (1993) and Dufwenberg and Kirchsteiger (2004) (DK), players reciprocate kind behavior, and kindness is measured by comparing the actual choice with the available alternatives. The most favorable choice for the other player has a kindness of 1, the least favorable a kindness of -1. In these models players only reward directly kind actions, i.e., kind actions to themselves. In the helping game, this is not excluded but it is rather unlikely that a high score of my recipient is the result of this recipient helping me. Nevertheless, the most unkind score is a score of 0, resulting in an kindness of -1, and the highest score results in a kindness of 1. Thus, scores below 50% are considered as unkind and scores above 50% as increasingly kind. Thus, we expect help only for scores above 50%, and we expect the helping probability to increase with the score. Since kindness is a relative concept in this model, i.e., since kindness is compared with the extremes of kindness, kindness is high for high scores even though it most likely results from help towards other people and actually had little impact on the donor who gets the possibility to reciprocate.

In the model of Levine (1998),[11] players differ in a parameter α_i that expresses how they value the other players' payoff. This parameter can vary from very altruistic, in which case they value the other players' payoff positively, to very spiteful, in which case they value the other players' payoff negatively. Furthermore, the actual evaluation of another player's payoff depends not only on one's own parameter α_i but also on the player's estimate of the other player's parameter α_j. This means that players are nicer to people who are nicer as measured by this parameter. Since all donors are confronted with the same

[11] Charness and Rabin (2002) have a model with a similar mechanism. Players reciprocate with "concern withdrawal" when they observe that the other player behaves selfishly or unkindly.

distribution of scores of recipients the donors with higher α_i are more likely to help, because they weight the recipient's payoff more positively. Because they help more, they get a higher score. This in turn implies that players with higher scores have higher α_i and deserve more help and hence donors with private score are more likely to help recipients with a higher score.

Consequently, the Levine model incorporates indirect reciprocity, since the score signals the α of the recipient, whereas in Rabin and DK, indirect reciprocity results because the score is a (relatively weak) signal of the kindness of the recipient towards the donor. In an infinitely large population, the score is not informative about the kindness towards the donor. Thus, in this case, indirect reciprocity is possible in the Levine model but it is not in the models of Rabin and DK.

3.4 Social Preferences and Reputation

Assume first that the players with public score all are selfish, but there is some form of pure indirect reciprocity, e.g. because there are players with private score who have social preferences. Following the same logic as Proposition 1 in Appendix 1, there are equilibria in which players with public scores behave indirectly reciprocally, i.e. they condition their helping on the recipient's score. These equilibria are characterized by the property that the players are indifferent between the potential scores, which implies that the expected payoff from having any of the scores that occur with positive probability should be equal.

If a higher score yields a higher probability to receive help, helping is less costly with a public score than with a private score. Thus, also non-selfish players are more likely to help when they have a public score. (Since selfish players do not help with a private score, it is obvious that they help more when they have a public score.) If there are non-selfish players, it is not necessary that all payoffs yield the same return. Suboptimally low scores are accepted by subjects who are willing to punish players for a low score, even when it is costly to them and suboptimally high scores are accepted by subjects who are willing to reward players for a high score, even when it is costly to them. Thus, it is likely that in an equilibrium with social preferences not all scores yield the same return. This implies that there is most likely an optimal score, and selfish players will try to keep their score on this level.

3.5 Hypotheses

In the introduction, we formulated three hypotheses, which can be derived from the analysis in this section. First, if there are subjects motivated by reciprocity, we will observe indirect reciprocity, i.e. the probability that donors help increases in the recipient's public score and in particular pure indirect reciprocity, i.e. this holds for donors with private score. Second, all models assume that subjects also care about their own payoff. Thus, we predict strategic reputation building, i.e., the average helping rate of donors with public score is higher than that of donors with private score. Third, strategic reputation building weakens the reciprocal relation, because when the score is public, concerns for the payoff will in some cases dominate reciprocal concerns.

4 Experimental Results

The overall experimental results are displayed in Figure 1, which shows the average helping behavior of donors with public and private scores for different public scores of the recipients.[12] 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 Figure 1 and Table 1:

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

Figure 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 score clearly increases with the score of the recipient, although the relation is monotone only for

[12] 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 a full score.

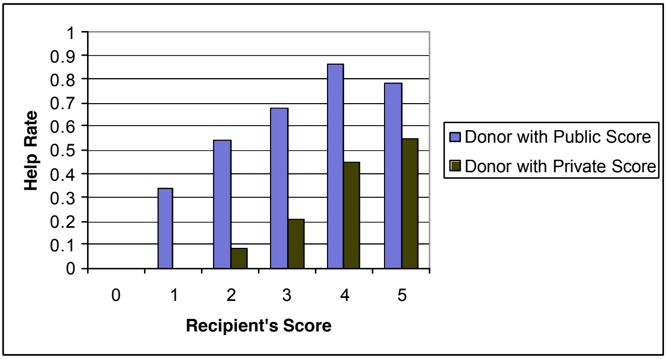


Figure 1: Donors' average help rate for recipients with full public score

donors with private score. 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 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 a 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.[13] This holds independently of whether we consider donors with public or private score.[14] That it does for donors with private score is evidence for pure indirect reciprocity.

One possible reason for pure indirectly reciprocal behavior is non-selfish preferences. If subjects care for other subjects, then there can be a motive to provide help even when there is no material benefit. How can

[13] Since the sessions are independent, the probability for an estimated positive correlation in all five sessions is (slightly smaller than) $\left(\frac{1}{2}\right)^5 = \frac{1}{32} < 5\%$. 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.

[14] 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.

	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%

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

the different models of non-selfish preferences account for the observed pattern? A concern for efficiency or inequality aversion in the sense of BO cannot explain that helping increases with the recipient's score. Both motives predict helping independent of the recipient's score.[15] If the players base the inequality assessment on the expected payoff, the FS model predicts that recipients with scores yielding higher payoffs will receive less help. Table 2 shows for all sessions the net profit that results from holding specific public scores, given the helping rates for these public scores. It shows that in all sessions the score of 4 was the most profitable and the average payoff increases monotonously for scores below 4, while Figure 1 shows that the helping rate increases weakly monotonously in the recipient's score. Thus, the prediction that recipients with scores yielding higher payoffs receive less help can be rejected. When comparing payoffs, advantageous inequality can also result from a high payoff of the donor. Therefore, if helping is based on inequality aversion, donors with higher payoffs should be more likely to help. It turns out that this prediction cannot be confirmed either. Regressions reveal even a negative relation between accumulated payoff and probability to help, also when controlling for time effects and individual fixed effects.

A better explanation for pure indirect reciprocity is provided by the reciprocity models. All these models predict a monotonous relation between score of the recipient and probability to get help. The models of Rabin (1993) and DK even go further and predict that there should only be helping for scores above 50%, i.e. for 3, 4, and 5. Interestingly, this is roughly the case. As Figure 1 reveals, for donors with private score there is no help for scores of 0 and 1 and less than 10% for a score of 2. So these models correctly capture the essential patterns of pure indirect reciprocity. As we argued above, however, Rabin

[15] As noted above, the recipient's score could be interpreted as a signal for the overall helping rate and hence the average payoff. The BO model then predicts a higher helping rate after observing a low score, the opposite of what we observe.

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 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

and DK capture indirect reciprocity only in a sense of stochastic direct reciprocity, based on the score as a noisy signal of previous behavior of the recipient towards the donor. Given our relatively large groups of subjects, this signal is indeed very weak. The model by Levine (1998) thus captures these results in a more convincing way. Levine's model is also consistent with a number of further observations. Since it allows for unconditional altruism, it can capture the relatively high rate of helping of donors with private score towards recipients with private score. More specific, if a recipient has a private score, the donor's estimate of the recipient's altruism should correspond to the average altruism in the reference group (even though they actually help less often). This is consistent with the helping rate towards recipients with private scores being similar to the average helping rate towards recipients with public scores.

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. As we will show later in the paper, this is, however, not the case.[16]

[16] 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

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

Result 2: *Donors with a public score help substantially more often than donors with a private score. 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 score is higher than of those with private score for any score of the recipient (including a private score, as can be seen from Table 1).[17] 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 score than for donors with private score for any recipient's 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 a panel data analysis (with the sessions as independent units of observations). We use a random-effects probit model.[18] The first model is

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

with $\Pr(Help)$ the probability that the donor helps, Φ the normal cumulative distribution function, $RsScore$ the recipient's score, $DPublic$ a dummy that takes the value 1 if the donor has a public score and 0 otherwise, and $DPublic \times RsScore$ the interaction effect between the recipient's score and the dummy for a donor with public score. α and β are significantly positive, providing further support for the first and second hypothesis, γ is significantly negative, supporting the third hypothesis (see (I) in Table 3).[19]

this period's recipient's score). This potential interpretation, however, appears to be valid only for donors with public score, because donors with private score do not have an incentive to find out what constitutes a successful score.

[17] The helping rate is strictly higher except for recipients with a score of 0, where the helping rate of both donors with public and private scores is 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 score is already 0 for recipients with a score of 1, this tie appears to simply result from censoring.

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

[19] We note that even if we restrict the analysis to donors with public score, the recipient's score still has a highly significant positive impact. Thus, although as predicted by the third hypothesis, the reciprocity of donors with a public score is reduced, it is not eliminated.

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 (since 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 suggest 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 maximal possible level.

As can be seen in Table 1, in each session for both recipients with public and private score the helping rates of donors with public score is about twice the helping rate of donors with private score. 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 score, 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 Figure 2 which shows the distribution (absolute frequencies on top of bars) of donors' full (public or private) scores. The mode of private scores is at a score of 0, with almost a uniform distribution over the remaining scores. For public scores, in contrast, the mode of the distribution is at a score of 4, with few cases of scores below 3 and hardly any below 2.[20] Interestingly, in all sessions the score that maximizes expected payoffs for the observed helping rates is 4 (see Table 2). Table 4 shows the result of a probit regression that supports the view that donors with public scores strategically maintain the optimal score of 4.[21] If their score falls

[20] Of the 19 full scores of 0, 15 come from the same subject, the only pure egoist. In all five sessions the mode for private scores is 0. For public scores the mode is 4 in three sessions. In one session the mode is 3 and in one session it is 5, with 4 being the second most frequent public score in both cases.

[21] We use individual fixed effects since we are interested in how individuals respond to changes in their score. If we do not use fixed effects, then we get a bias, because players who always have a score of 5 trivially help more than those who have a score of 5 only occasionally.

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

Table 3: Random-effects probit model for the help choice. (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 Criterio. Data is restricted to the cases with full score for both players. *: $p < 0.05$, **:$p < 0.01$, ***:$p < 0.001$.

below 4, they increase their probability to help, if it is greater than 4, they decrease it. This is only true for donors with a public score. The helping rate of donors with private scores does not depend on whether their score is larger or smaller than 4.[22]

While these results show that donors are clearly influenced by strategic considerations, they also exhibit pure indirect reciprocity, not driven by strategic concerns.[23] As shown in Proposition 1 in Appendix 1,

[22] Since a private score cannot be observed, it cannot yield any direct benefits (only possibly indirect spill-over effects from raising average cooperativeness in the session). Since maintaining a positive score is, however, costly, the optimal private score is 0.

[23] To some degree, this has to be the case. Since all our subjects have a public score in one half of the experiment and

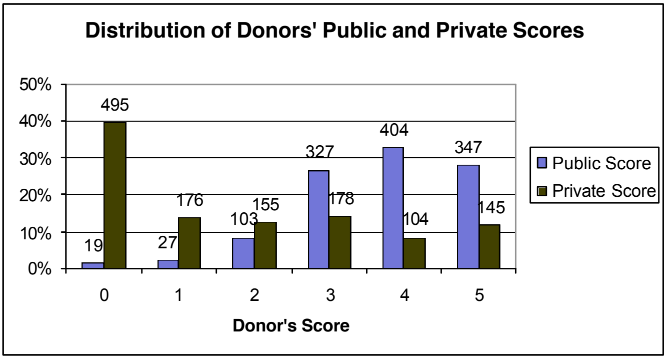


Figure 2: Distribution of public and private (post-decision) donors' scores for all interactions where the donor had a full score (except for following donors' decisions in the last period because in that case the resulting score could not possibly be relevant for future interaction). Absolute numbers 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 a private score were chosen slightly more often as donors).

purely selfish subjects would only differentiate their helping if they were indifferent between the scores that occur with positive probability. Table 2 shows that 4 was the optimal score in all 5 sessions. If the *expected* payoff was the same for the highest three scores, the probability that in all five session the *observed* maximum is at the same score equals $3 \cdot \left(\frac{1}{3}\right)^5 = 0.012$. So, the hypothesis that there is no systematic difference in the returns to different scores can be rejected.

Players who are motivated by reciprocity concerns would be more willing to deviate from the optimal score in order to reward or punish others. Hence their public scores should vary more across time than that of a selfish subject. On the other hand, they should help more often when they have a private score than

a private score in the other half of the experiment, the subjects with public and those with private score are the same participants. The motivation behind their reciprocity when they have a private score should also be present when they have a public score. The question is hence more precisely whether the strategic considerations completely dominate the reciprocal concerns. Our results show that this is not the case.

	Donors with public scores	Donors with private scores
$RsScore$	0.5614***	1.0669***
	(0.0893)	(0.1427)
$ownscore < 4$	0.7004***	-0.0611
	(0.1854)	(0.4204)
$ownscore > 4$	-0.8339***	-0.0542
	(0.2374)	(0.4949)
$Indiv.\ fixed\ effects$	yes	yes
$const$	-2.4768***	-4.3056***
	(0.4542)	(0.07080)

Table 4: Probit models for the help choice of donors with public and with private scores. RsScore is the recipient's score. "ownscore < 4" is a dummy that equals 1 when the donor's own score is smaller than 4. "ownscore > 4" is a dummy that equals 1 when the donor's own score is greater than 4. Individual fixed effect are used in both regressions. Robust standard errors take the dependence of the data within sessions into account. Only data with full scores for donor and recipient is used.

the selfish subjects. This yields the prediction that the helping rate when a subject has a private score is correlated with the variance of the subject's own score when the subject has a public score. This is indeed the case. In all sessions there is a positive rank correlation between the two, which occurs at random with a probability of $\frac{1}{32}$.[24] We further observe:

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

An advantage of our design is that we can study the importance of strategic reputation building on an individual basis by comparing the helping rates with public and with private score within subjects. Table 5 shows a classification of subjects. We call a subject strategic if her helping rates are generally higher in the part of the experiment where she has a public score than in the part where she has a private score. A strategic subject is called strongly strategic if the helping rates with public score are in most cases

[24] Actually, in all but one seesion the correlation was highly significant, but this is not a valid test since the observations are not independent.

	Pure Strat	Strong* Str.	Weak Str.	Non-Str.	Total
Reciprocal	8	12	14	4 (5)	38 (39)
Non-Reciprocal	12	11	7	6 (11)	36 (41)
Simple Egoist				1	1
Simple Altruist				4	4
Negativ Rec. Altr.				1	1
Total	20	23	21	16	80

Table 5: Classification of individual subjects (in absolute numbers). Numbers in parantheses include special types of reciprocal (negatively reciprocal altruist) and non-reciprocal (simple egoist and simple altruist) non-strategic types. These types are listed separately in the third to fifth rows. Strong* Strategic (third column) refers to the players who are strongly strategic but not purely strategic.

at least twice the helping rate with private score.[25] Otherwise she is called weakly strategic. Finally, a pure strategist never helps when she has a private score, but does so several times otherwise.[26] There are several special cases of non-strategic subjects. Simple egoists never help, simple altruists always help and negatively reciprocal altruists always help if the recipient has a private score or if the recipient's public score is above some cut-off level, but not for a lower public score. A subject is classified as reciprocal if there is a clear positive relation between the recipient's score and the helping rate.[27]

[25] For the classification of both strategic and strongly strategic, we allowed deviations from the respective criterion for one value of the recipient's public score and in that case required it to hold strictly for at least two values of the recipient's score. We always required the respective criterion to hold for the interactions where the recipient had a private score because their number was much higher than the interaction with recipients of any single public score, so a violation of the criterion could not be driven by a small number of observations.

[26] Note that the pure strategists are a subset of the strongly strategic subjects, but in the table the third column shows those stronly strategic subjects which are not purely strategic.

[27] We allowed one exception from the criterion in the sense that for one low score the helping rate was allowed to be higher than for one or several higher scores or for one high score the helping rate was allowed to be lower than for one or several lower scores. In these cases we required at least two either low scores where the helping rate was lower than that for all higher scores or high scores where the helping rate was higher than for each lower score. A flat helping rate in case the donor had a private score was allowed if the helping rate in case he had a public score showed a clear positive relation. For most subjects, the classification was straightforward, because there was either a clear monotone relation or none at all.

	Strongly Strategic	Weakly or Non-Strategic	Total
Reciprocal	1.14 (20)	0.69 (19)	0.92 (39)
Non-Reciprocal	1.23 (23)	0.87 (18)	1.08 (41)
Total	1.19 (43)	0.78 (37)	

Table 6: Payoffs relative to average session payoff for strongly strategic versus weakly or non-strategic and for reciprocal versus non-reciprocal players, number of players in the respective category in parantheses.

As Table 5 shows, the majority of subjects is clearly strategic. The crucial qualitative aspects of the distribution of types (i.e. that only a minority is non-strategic and that the share of reciprocal and non-reciprocal subjects is about equal) also hold for the two subsets of subjects that have a public score in the first or in the second half of the experiment, as well as in each session.[28] Surprisingly, there is only one negatively reciprocal altruist, which intuitively appears to be a perfectly reasonable and in particular socially desirable type (helps in general but punishes egoists). Some of the 4 simple altruists might be negatively reciprocal altruists, because they never encountered a recipient with a score below 2. Thus, we find a high share of subjects who are reciprocal but even a higher share that are strategic. Interestingly, 40% of the pure strategists and 52% of the other strong strategist 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.[29] 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. We next observe:

Result 4: *Strongly strategic subjects obtain significantly higher payoffs than subjects who are at most weakly strategic. Reciprocal subjects obtain lower payoffs than non-reciprocal subjects.*

Confirming straightforward intuition, strategic reputation building pays, whereas reciprocity does not. Table 6 shows the average payoffs (relative to the average payoffs in the session) of subjects by being

[28] The number of reciprocal subjects varies between 6 and 10 among the sessions, and the number of non-strategic subjects between 2 and 5.

[29] This can be seen as being strategic on a higher level, because due to the matching procedure, donors could profit from inducing others to help, either by later being matched with them again or by indirect effects.

reciprocal and strongly strategic, where we pool the pure and other strongly strategic players on the one had and those that were classified as weakly or non-strategic on the other hand. Clearly, the former outperform the latter, which does not come as a surprise because being strategic implies, conditioned on the public score, a lower private score and hence lower costs for helping. The advantage of the strongly strategic players is, however, remarkably large.[30] More importantly, it pays not to be reciprocal, apparently because being reciprocal distracts from perfectly fine-tuning one's own score (or, in case of private scores, is a pure waste).[31] 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 the cooperation. Given that the relative payoff of the non-reciprocal strongly strategic players is almost twice that of the reciprocal weakly or non-strategic players, the evolutionary process would be quite fast for any sufficiently payoff-sensitive dynamic. Finally, we look at the time dependence of our data and observe:

Results 5: *End-game effects are consistent with the major patterns of behavior.*

Figure 3 shows the development of the average helping rates in the first 40 and the second 40 periods. While there is a clear drop in the last two periods in both cases, the helping rate is remarkably stable until the third to last. Since the expected return to a high score decreases sharply towards the end of the experiment, one might have expected helping rates to drop earlier. An analysis of the sources of the end-game effect is remarkably consistent with our above classification of subjects into purely strategic, strongly strategic, weakly strategic and non-strategic. A subject who helps primarily in order to strategically build a score would be expected to lower his or her helping rate in the last periods when having a public score.

[30] If we study the data in a more disaggregated way, we find that the payoff for the purely strategic is slightly higher than that for the other strongly strategic and the payoff for the weakly strategic is substantially higher than for the non-strategic. Since the numbers of observations is too low for some categories in some sessions to derive meaningful results and since the largest difference is between strongly (but not purely) strategic and weakly strategic we aggregated the data in two categories for the present analysis.

[31] The average relative payoffs are larger for the strongly strategic than for the at most weakly strategic in all 5 sessions (for the non-reciprocal, for the reciprocal as well as for the whole sample) and hence we can reject the hypothesis that the strongly strategic do not do better at $p = 5\%$. The non-reciprocal do better than the reciprocal in only four sessions and hence this test misses statistical significance.

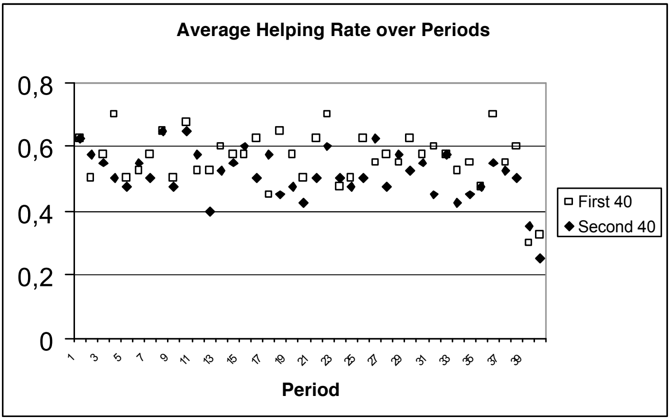


Figure 3: Average helping rates over all sessions for the respective 40 periods of the first and the second half of the experiment

The comparison of individual subjects' helping rates in the last two periods with their overall helping rates when they have a public score is consistent with this expectation. Out of 17 subjects whom we classified as purely strategic and who have been a donor at least once in the last two periods, only 2 increase their helping rate, while 15 lower it.[32] The corresponding numbers for the other strongly strategic players are 4 and 14, for the weakly strategic they are 5 and 8 and for the non-strategic they are 4 (plus 4 with a constant helping rate) and 5. Thus, the end-game behavior clearly corresponds to our classification of subjects in terms of strategic behavior. Subjects classified as strongly strategic exhibit a clear drop in helping behavior towards the end while those classified as weakly or non-strategic do not.[33]

Furthermore, the end-game effect is almost exclusively restricted to the subjects with a public score,

[32] Most subjects were a donor only once in the last two periods. For these players, increasing the helping rate means helping this one time and decreasing the helping rate means not helping this one time.

[33] Overall helping rates do not differ dramatically between the different categories in the phase when subjects have a public score. For purely strategic subjects it is 67%, for the other strongly strategic 68%, for weakly strategic 84%, and for non-strategic 72%. Therefore, the observation concerning end-game effects is not an artifact of differences in overall helping rates.

consistent with our interpretation that a substantial share of helping behavior by donors with a public score is driven by strategic reputation building, while that of subjects with a private score is pure indirect reciprocity (and not, for example, strategic in a sense of trying to provide incentives for others to help). For players with a private score, the helping rate in the last two periods of the first phase is nearly equal to the average rate (period 39: 25%, period 40: 35%, overall average: 38%, the rate is below 25% already in three earlier periods). In the second phase the helping rate is also only slightly below the average (period 39: 28%, period 40: 19%, overall average: 31%, the rate is below 19% already in six earlier periods). In contrast, for players with a public score, the helpings rate drops dramatically below the average in the last two periods of both the first and the second phase (First phase: period 39: 33%, period 40: 29%, overall average: 74%, the helping rate is above 50% in all previous periods; second phase: period 39: 41%, period 40: 29%, overall average: 72%, the helping rate is above 45% in all previous periods). In particular, the helping rate of donors with a public score almost drops to the level of donors with a private score, which we would expect if the difference in their behavior is driven by strategic reputation building that cannot matter in the last period.

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 to the helping experiment by Seinen and Schram (2006). In their experiment, all subjects could build up an image score (or none in the control treatment) 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 non-selfish cooperative behavior (helping by donors with private scores) and strategic cooperative behavior (the difference in behavior of donors with private and public scores). The average helping rate of donors with private score of 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 a public score is evidence for strategic reputation building. From a more specific perspective, we are the first to find clear evidence for indirect reciprocity even in the absence of strategic incentives for reputation building, but we also find very strong effects of strategic reputation building. Specifically, 80% of subjects react to strategic incentives, including more than 50% whose helping rates more than double and 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 allows to shed 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 the cooperation by exploiting the system are clearly present and more successful. Put differently, the argument by Leimar and Hammerstein that the set of potential types chosen in the simulations by Nowak and Sigmund is too restricted is not only valid 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. Furthermore, the helping rate of 37% by donors with a private score contradicts the evolutionary model by Nowak and Sigmund. In their model indirect reciprocity evolves where players can build a reputation. The donors with a private score, however, cannot build a reputation. Their helping behavior would be consistent with the Nowak and Sigmund approach only if one assumes that they behave maladaptively in this environment. On the other hand, the subjects behave very adaptively, because donors with a public score help twice as often and hence seem to clearly understand the incentives of reputation building. Hence, there appear to be further underlying motivations.

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

[1] Alexander, Richard D., 1987. *The Biology of Moral Systems*. New York: Aldine de Gruyter.

[2] Andreoni, James and Petrie, Ragan, 2004. "Public Goods Experiments without Confidentiality: a Glimpse into Fund-Raising." *Journal of Public Economics* 88, 1605–1623.

[3] Berg, Joyce, Dickhaut, John, and McCabe, Kevin, 1995. "Trust, Reciprocity and Social History." *Games and Economic Behavior* 10, 122–142.

[4] Bolton, Gary E., Katok, Elena, and Ockenfels, Axel, 2004. "How Effective Are Online Reputation Mechanisms? – An Experimental Study." *Management Science* 50, 1587-1602.

[5] Bolton, Gary E. and Ockenfels, Axel, 2000. "ERC – A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90, 166–193.

[6] Charness, Gary and Rabin, Matthew, 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics* 117, 817–869.

[7] Dufwenberg, Martin and Kirchsteiger, Georg, 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior* 47, 268–298.

[8] Falk, Armin, and Fischbacher, Urs, 2006. "A Theory of Reciprocity." *Games and Economic Behavior* 54, 293-315.

[9] Fehr, Ernst, Kirchsteiger, Georg, and Riedl, Arno, 1993. "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics* 108, 437–460.

[10] Fehr, Ernst and Schmidt, Klaus M., 1999 "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114, 817–868.

[11] Fischbacher, Urs, 2007. "Z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10, 171-178.

[12] Harbaugh, William T., 1998. "The Prestige Motive for Making Charitable Transfers." *American Economic Review* 88, 277–282.

[13] Hagen, Edward H. and Hammerstein, Peter, 2006. "Game Theory and Human Evolution: A Critique of some Recent Interpretations of Experimental Games." *Theoretical Population Biology* 69, 339–348.

[14] Keser, Claudia, 2002. "Trust and Reputation Building in e-Commerce" *Working paper,* IBM T. J. Watson Research Center.

[15] Leimar, Olof and Hammerstein, Peter, 2001. "Evolution of Cooperation through Indirect Reciprocity." *Proceedings Royal Society London: Biological Sciences* 268, 745–753.

[16] Levine, David K., 1998 "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1, 593–622.

[17] Milinski, Manfred, Semmann, Dirk, and Krambeck, Hans-Jürgen, 2002. "Donors to Charity Gain both Indirect Reciprocity and Political Reputation." *Proceedings Royal Society London: Biological Sciences* 269, 881–883.

[18] Nowak, Martin A. and Sigmund, Karl, 1998. "Evolution of Indirect Reciprocity by Image Scoring." *Nature* 393, 573–577.

[19] Rabin, Matthew, 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83, 1281–1302.

[20] Seinen, Ingrid and Schram, Arthur, 2006. "Social Status and Group Norms: Indirect Reciprocity in a Helping Experiment." *European Economic Review* 50, 581-602.

[21] Wedekind, Claus and Milinski, Manfred, 2000. "Cooperation through Image Scoring in Humans." *Science* 288, 850–852.

Appendix 1

In this appendix, we analyze the infinitely repeated game. To keep the analysis tractable, we make a number of simplifying assumptions. We assume that the game has the following structure: Half of the players can help in the even rounds and the other half of the players can help in the odd rounds. Furthermore, we assume that players discount their future payoffs after a period in which they are recipients. The discount factor equals δ. We also slightly modify the way in which the score is maintained. In the experiment, the score consists of the last 5 periods. For the analysis of the infinitely repeated game, we assume that the score consist of H decisions of the past. Whenever the score is updated, the new decision does not replace the oldest decision in the score but one randomly selected decision. To permit the player to condition his decision on his own score, he gets to know which score has been deleted before he has to decide whether to help or not. This way, it is for instance possible that the player exactly keeps his score constant.

We assume that the probability that a player is paired with a player with whom he directly or indirectly interacted before equals zero. (This is a good approximation if there are many players.) Thus, players with private scores have no incentive to help. If M out of N players do have a public score, then recipients with public score meet a player with public score with a probability of $\frac{M-1}{N-1}$. Consequently, every equilibrium in a game that also contains players with private scores corresponds to an equilibrium in which there are no players with private score but the benefit equals only $b\frac{M-1}{N-1}$. For this reason, we consider a game in which there are only players with public score for the remainder of this section.

Let h_k be the probability of getting help as a recipient with a score of $k \in \{0, ...H\}$, let p_k be the probability of having a score of k as a recipient and q_k be the probability of having a score of k as a donor. In the latter case, k ranges from 0 to $H-1$ because the decision as donor will be the H^{th} element in the score (since one element of the score is deleted before a donor makes a decision).

The basic idea behind the characterization of the equilibria consist in the observation that players will condition their choice on the recipient's score only if the future benefit from the higher score exactly offsets the immediate cost of helping, i.e., the player has to be indifferent between helping and not helping. As will be shown in the following proposition, this is the case if the difference between the helping probabilities of the adjacent scores equals $\varphi := \frac{c}{Hb}(1 + (1-\delta)(H-1))$.

In the following proposition, we will show that there are infinitely many equilibria if $H\varphi < 1$ (see a) and we will present some properties of all equilibria (independent of whether $H\varphi < 1$ or not; see b).

Proposition 1 *(a) If $H\varphi < 1$ then for all $\gamma \in [0,1 - H\varphi]$, there is an equilibrium, in which $h_k = \gamma + k\varphi$. In this equilibrium, the probability p_k of having a score k is positive for all scores k if $\gamma \in (0,1 - H\varphi)$.*

(b) Consider a stationary, symmetric, subgame perfect equilibrium. If $p_k > 0$ for $k \in I = \{\underline{k}, \underline{k}+1, ..., \bar{k}\}$, then there is a value γ such that $h_k = \gamma + k\varphi$ for $k \in I$.

Remarks: Note, that $\varphi \to \frac{c}{bH}$ for $\delta \to 1$. This implies that for the parameters of the experiment for sufficiently high δ, $\varphi \approx \frac{6 \cdot 15}{15 \cdot 7 \cdot H} < \frac{1}{H}$, which means that the difference between the helping rates for the highest (H) and lowest (0) scores is smaller than 1. Thus, in this case, Proposition 1(a) applies and there are many equilibria in which all scores occur. Furthermore, there are equilibria in which there is no help. Obviously, there is an equilibrium with $h_k = 0$ for all k. Additionally, helping rates $h_k \leq k\varphi$ yield an equilibrium in which the probability of having a score of zero equals 1. With $\gamma = 1 - H\varphi$, there is also an equilibrium in which the helping probability equals 1.

Interestingly there are many equilibria. In particular, there are also equilibria in which the donors condition their helping on their own score. For an example of an equilibrium in which not all scores occur with positive probability, we use the parameters from the experiment and take $\delta \to 1$. In this case donors are indifferent between adjacent scores if the helping rate differs by $\frac{6}{35}$. We present an equilibrium in which players maintain a score between 2 and 3. In this equilibrium only donors with a score of 2 condition their help on the score of the recipient. These donors help recipients with a score of 2 with probability $\frac{3}{5}$, and help recipients with a score of 3 with probability $\frac{31}{35}$; in the other cases the recipients does not get help. If the donor has a score of 1, the recipient gets help for sure and in all other cases, the recipient does not get help. In steady state, there are 16% recipients with a score of 2 and 84% recipients with a score of 3. Recipients with score 0, 1, and 5 get help with probability 0.064, recipients with score of 2 get help with probability of $0.424 > 2 * \frac{6}{35} + 0.064$ and recipients with a score of 3 get help with probability $0.424 + \frac{6}{35}$. In this equilibrium donors are indifferent of having a score of 2 or 3, but strictly prefer these scores to all other scores.

It is also possible that the support of the probability distribution of p_k is not connected. In that case, players are indifferent between the scores within each connected component, but not necessarily between scores of different connected components. However, the utility differences *between components* cannot be too large, because otherwise the players with scores in one connected component have an incentive to "move" to the scores in the other component. The maximum possible utility difference depends on the positions of the connected components and on the discount factor. For $\delta \rightarrow 1$, the maximum possible difference goes to zero. Part (b) of the proposition thus implies that the graph of h_k lies on parallel lines on the different connected components of the support of the probability distribution of p_k. It lies on one line when $\delta \rightarrow 1$.

Proof. We start with the proof of (b). Let V_k, $k \in \{0, 1, ..., H-1\}$ be a donor's expected profit before his helping decision if he has a score of k, i.e. after one of his decisions is cleared from the score. First, we show that for a score $k \in \{\underline{k}, ..., \bar{k}-1\}$, the donor is indifferent between helping and not helping. If this were not the case then, because of the symmetry, all donors would either help or not help when they have score k. Consider the case in which the donors always help. (For the case in which all donors do not help, a similar argument applies.) Because of the stationarity condition the distribution of scores has to be constant over time. We apply this argument to the distribution of the scores of the donors. It implies in particular, that the probability of a donor having a score of at least k cannot change. Thus, the donors with score $k-1$ never help because the donors with score k and higher will also have a score of at least k when they decide the next time. But if the donors with score $k-1$ never help and the donors with score k always help, the score k will never occur for the recipients. This contradicts the assumption $p_k > 0$.

If the recipient scores between $\underline{k}$ and $\bar{k}$ occur with positive probability, then the donor scores between $\underline{k}-1$ and $\bar{k}$ occur. For k between $\underline{k}-1$ and $\bar{k}-1$ we get

$$V_k = -c + bh_{k+1} + \delta(\frac{k+1}{H}V_k + \frac{H-k-1}{H}V_{k+1}) \tag{1}$$

and for k between $\underline{k}$ and $\bar{k}$,we get

$$V_k = bh_k + \delta(\frac{k}{H}V_{k-1} + \frac{H-k}{H}V_k) \tag{2}$$

We can use the same V_k in both formulas, because for k between $\underline{k}$ and $\bar{k}-1$, the donor is indifferent. Using (1) for $k-1$ and combining it with (2), we get $V_k = V_{k-1} + c$ for k between $\underline{k}$ and $\bar{k}$.

Let us define $k_0 = \underline{k} - 1$ and $d = k - \underline{k} + 1$. Then, we get from (2)

$$\begin{aligned}
V_{k_0+d} &= bh_{k_0+d} + \delta(\frac{k}{H}V_{k_0+d-1} + \frac{H-k}{H}V_{k_0+d}) \\
V_{k_0} + dc &= bh_{k_0+d} + \delta(V_{k_0} + \frac{Hd - k_0 - d}{H}c) \\
V_{k_0}(1-\delta) &= bh_{k_0+d} - c(d - \delta\frac{(H-1)d - k_0}{H}) \\
&= bh_{k_0+d} - \frac{c\delta k_0}{H} - \frac{cd}{H}(H - \delta(H-1))
\end{aligned}$$

This is true for any d, in particular for $d = 0$. Hence, for all d:

$$V_{k_0}(1-\delta) = bh_{k_0+d} - \frac{c\delta k_0}{H} - \frac{cd}{H}(H - \delta(H-1)) = bh_{k_0} - \frac{c\delta k_0}{H}$$

Thus,

$$\begin{aligned}
h_k &= h_{k_0+d} = h_{k_0} + d\frac{c}{Hb}(H - \delta(H-1)) \\
&= h_{k_0} + (k - k_0)\frac{c}{Hb}(1 + (1-\delta)(H-1)) \\
&= h_{k_0} - \varphi k_0 + k\varphi
\end{aligned}$$

This proves (b).

Let us now prove (a) for $\gamma \in [0, 1 - H\varphi]$. As shown above, players are indifferent in their helping decisions if $h_k = \gamma + k\varphi$. We have described above the equilibria for $\gamma = 0$ and $\gamma = 1 - H\varphi$. Let us now consider $\gamma \in (0, 1 - H\varphi)$. We will show that there are probabilities p_k such that the helping probabilities h_k (unconditional on the donor's score) generate a stationary process. In the stationary equilibrium, players always face the same distribution of scores of the recipients. Therefore, we can define the a-priory probability that a donor will help, i.e. the probability that the donor helps before he knows the score of the recipient. We denote this probability by α. Then, the score follows a Markov process which is determined by α and the score forgetting process. It can be shown that the eigenspace of the eigenvalue 1 for the corresponding Markov matrix has dimension 1. Thus, there is a *unique* stationary distribution of scores for the recipients $p_k(\alpha)$ for every α. Now, in equilibrium a recipient with score k gets help with probability $h_k = \gamma + k\varphi$. With the distribution given by α, we are perhaps not yet in equilibrium, but we can calculate the unconditional probability $\rho(\alpha)$ that a recipient gets help if with score k he is helped with probability $h_k = \gamma + k\varphi$ and the distribution of scores is given by $p_k(\alpha)$. It equals $\rho(\alpha) = \sum_k p_k(\alpha)h_k$. The function ρ is continuous in α. Further, $\rho(0) = \gamma > 0$, and since $\gamma < 1 - H\varphi$, $\rho(1) = \gamma + H\varphi < 1$ holds. Thus, there is some α^* with $\rho(\alpha^*) = \alpha^*$. This provides the required equilibrium. In this equilibrium all scores occur with positive probability, because $\alpha \in (0, 1)$ and because helping does not depend on the donor's own score. ■

While we have simplified the environment, in particular by assuming that the game is repeated infinitely often, the basic insights intuitively apply to the setting of our experiment as well in case players disregard the final end. In particular, in an equilibrium among completely selfish players, scores other than 0 should only occur if adjacent scores that appear with positive probability yield the same expected payoff.

Appendix: Instructions
(Original Instructions were in German)

General Instructions

You are taking part in an economic experiment, which is being financed by various research promoting foundations. If you read the following instructions carefully, you can - depending on the decisions you will make - influence your own earnings as well as the earnings of the other participants of this experiment. It is, therefore, important that you pay attention to the instructions given below.

The instructions distributed are intended for your personal information only. **Absolutely no communication whatsoever is allowed for the duration of the experiment.** Please address questions you might have to us directly. Violation of this rule leads to the exclusion both from the experiment itself and from all pertaining payments.

The experiment is divided into **periods**. During this experiment we do not deal with francs, but with points. Your income from each period will, therefore, be calculated in points. The total amount of points achieved in the course of the experiment will be converted into francs at the rate of

1 point equals 10 rappen [100 rappen = 1 Swiss Franc].

At the beginning of the experiment you are allotted an endowment of 100 points, thus representing 10 francs.

In each period you form a group with **one** other participant. These groups of two are in each period newly formed at random. It is possible, though not probable, that you will be linked with the same participant in two consecutive periods. You cannot recognize the other participants, and hence do not know whether you have been in a group together with the current other participant before. This guarantees the anonymity of your decision.

Each group consists of one participant with the **part A** and one participant with the **part B**. Both parts are, in each period, randomly and independently assigned. The probability of being assigned part A for a period is 50 %, irrespective of the part held in the previous period. Therefore, it is possible that you will assume part A or part B in several consecutive periods.

Specific Instructions for the Experiment's Procedure

Decisions to be made by the participants

During each period, in which you assume part A, you determine whether or not you want to help the other participant of your group (who holds part B). If you assume part B no decision is required from you. If you, as the holder of part A, decide to help the other participant of your group, you will be charged with a cost of 6 points, and the other participant of your group is given 15 points. If you decide not to help the other participant of your group, you suffer no cost, and the other participant receives nothing, resulting, for both of you, in the same amount of points as at the beginning of the period.

Participants' information types

The participants differ from each other insofar as other participants are, or are not, informed of the decisions made. Participants, whose decisions are communicated to the other participants, are referred to as **Info types**. The experiment comprises two stages consisting of 40 periods each. At stage one, i.e. during the first 40 periods, one half of the participants are info types. At stage two, the other half of the participants become info types. Thus, you will, like all other participants, be an info type **either** during the **first** 40 periods **or** during the **second** 40 periods. You will always know if you are an info type or not. If during the first 40 periods you were an info type, we will inform you at the end of these 40 periods that for the rest of the experiment you will no longer be an info type and vice-versa. Regardless whether you are an info type or not, you can in each period be matched both with another info type or to a non-info type.

Information on info types

The **last five decisions** made by the info types are being computer-saved, i.e. saved will be the number of times an info type (with part A) granted help and the number of times he denied help. When an info type then assumes part B, this information is given to the other participant of the group (assuming part A). This means that the participant with part A learns how many times the participant with part B granted help during the last five periods and how many times he did not. If at this stage the participant with part B assumed part A in less than five periods, the participant with part A is informed of decisions B made in these periods.

If a participant is not an info type, no information on his decision-making is saved. In particular this means that no-one is informed about the decisions made at the stage where one is not an info type. Thus if at stage two of the experiment you are an info type, no information on the decisions you made at stage one will be passed on to another participant.

Participants with part B are given **no** information on participants with part A.

If you are an info type, whose current decisions as participant with part A are passed on to later participants assuming part A, you are, at the beginning of a period, informed of how you decided during the last five periods with part A (or during less than these five periods if you assumed part A less than five times). This information is submitted to you regardless of which part, A or B, you assume.

Stage two of the experiment

On completion of the 40 periods of stage one and after a short break we will get started with stage two, again consisting of 40 periods. The info types of stage one are no longer info types, and the non-info types of stage one become the info types of stage two. At stage two, all information on the decisions made at stage one are no longer available. This means that the number of periods with part A about which information is released, starts at zero for all participants.

However, the amount of points earned at stage one are carried over to stage two.

Procedure on the Computer

The screen shown to both participants is divided in two sections. The **upper section of the screen** is independent on whether you assume part A or part B.

Information given in the upper section of the screen

Each period reveals, in the **upper section of the screen,** the part you assumed in the previous period as well as the decision the participant with part A made in the last period (see figures 1 and 2 below). Furthermore, you are shown your actual balance of points. As an info type you will also see how many times during the last five periods as A (or during all previous periods as A, if they amount to less than five) you granted help to the participant with part B and how many times you denied it (see the example in figure 1). This is for your information. In the example in Figure 1 you have been the participant with part A during the last period, granting help to the participant with part B. During the last five periods with part A, you granted help twice and denied it three times. The current balance is 121 points. The example in Figure 2 shows the upper section of the screen, if you are not an info type. During the last period you assumed part B and were granted help. Your current balance is 121 points.

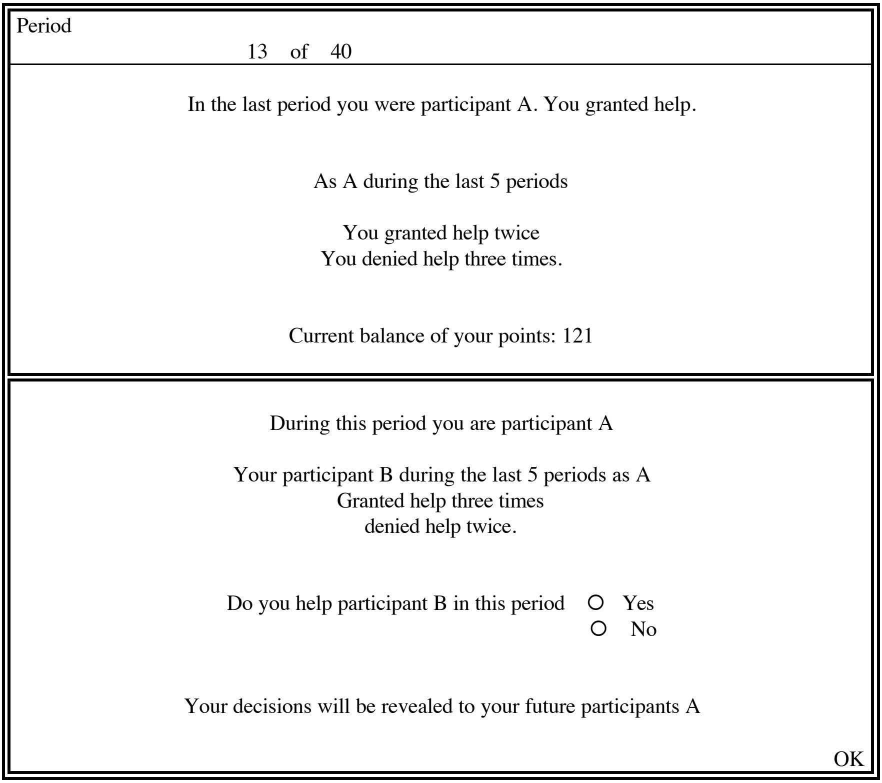
Decision-making section for participants A

If you are the participant with part A, you make your decision in the **lower section of the screen**. If the other participant of your group, i.e. the participant with part B, is an info type, you are informed about B's last five decisions (i.e. about the last five periods where he assumed part A). In the event that the other participant of your group, i.e. the participant with part B, is no info type, you are informed about the fact that no information is released to you. The screen below shows that the participant with part B granted help three times and denied it twice during the last five periods where he assumed the part of A.

Below you will see the following question: '**Do you help participant B in this period?**" beside the two fields "Yes" and "No". Mouse-click one of these fields and activate the "OK" button. **If you choose "Yes" your balance of points will be reduced by 6 points and participant B's balance will be increased by 15 points. If you choose "No" neither your nor participant B's balance will be changed.**

Besides, you will learn if you are an info type, which in this example applies. Thus your decision will in future periods, where you assume part B, be revealed to the participant with part A as long as these decisions belong to your last five decisions as the participant with part A and as long as you are at the same stage of the experiment.

Figure 1: Screen for participants with part A



Period

13 of 40

In the last period you were participant A. You granted help.

As A during the last 5 periods

You granted help twice
You denied help three times.

Current balance of your points: 121

During this period you are participant A

Your participant B during the last 5 periods as A
Granted help three times
denied help twice.

Do you help participant B in this period ○ Yes
○ No

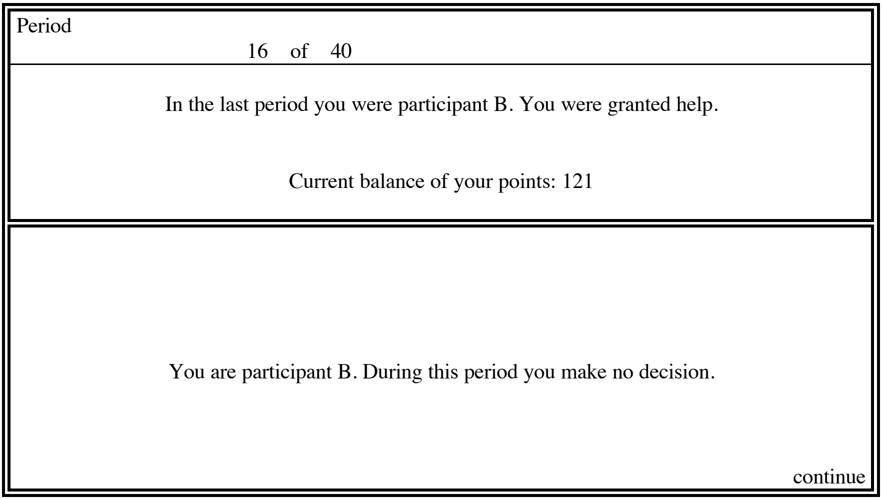
Your decisions will be revealed to your future participants A

OK

Lower section of the screen for participants with part B

The lower part of the screen only informs you that during this period you are not to make any decision.

Figure 2: Screen for participants with part B



Period

16 of 40

In the last period you were participant B. You were granted help.

Current balance of your points: 121

You are participant B. During this period you make no decision.

Control Questionnaire

Please answer all questions. Wrong answers have no consequences whatsoever! Address any questions to us!

1. Participant A has 121 points, participant B has 112 points. Participant A helps participant B. The balance of points of the participants is:

participant A: ....

participant B: ....

2. Participant A has 145 points, participant B has 127 points. Participant A denies participant B help. The balance of points of the participants is:

participant A: ....

participant B: ....

3. Suppose you are an info type. During the last five periods you made the following decisions: "help denied", "help denied", "help granted", "help granted", and "help denied" (in this sequence). You are now again A. In the event that you now help and that in the next period you assume part B: which information on your decisions will be released to participant A?

you granted help .... times

you denied help .... times

4. Suppose that during the first stage of the experiment you are an info type. In how many periods, at the most, is the decision you make in period 37 revealed to another participant?

5. Suppose you had the part of B three consecutive times. What is the probability of you again assuming part B during the next period?

THURGAU INSTITUTE
OF ECONOMICS
at the University of Konstanz

Hauptstr. 90
CH-8280 Kreuzlingen 2

Telefon: +41 (0)71 677 05 10
Telefax: +41 (0)71 677 05 11

info@twi-kreuzlingen.ch
www.twi-kreuzlingen.ch