Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game[¤]

Dirk Engelmann[y] and Urs Fischbacher[z]

September 15, 2004

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 …nd that while pure indirect reciprocity appears to be important, the helping choice seems to be in‡uenced at least as much by strategic considerations. 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 Classi…cation: C92

Keywords: indirect reciprocity, reputation, experimental economics

[¤]We thank James Cox, Alan Durell, Ernst Fehr, Simon Gächter, Werner Güth, Ste¤en Huck, Michael Kosfeld, Manfred Milinski, Wieland Müller, Ronald Oaxaca, Andreas Ortmann, Arno Riedl, Rupert Sausgruber, Dirk Semmann, Georg Weizsäcker, and Viatcheslav Vinogradov 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 …nancial support from the Deutsche Forschungsgemeinschaft (DFG, grant No. EN 459/1).

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

[z]University of Zurich, Institute for Empirical Research in Economics, Bluemlisalpstr. 10, 8006 Zurich, Switzerland, …ba@iew.unizh.ch.

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, Falk and Fischbacher, 1999, Dufwenberg and Kirchsteiger, 2004) 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 a¤ected 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 di¢cult. So, 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.[1]

The term "indirect reciprocity" was introduced by Alexander (1987). He argues that individuals' behavior towards others is not only in‡uenced by their own experience but also by their observations of other's behavior. According to Alexander, indirect reciprocity works through reputation and status and provides the evolutionary basis for moral systems prescribing cooperation.[2]

Harbaugh (1998) argues (and provides supporting …eld data) that donations to charity are in part driven by a prestige motive. The bene…t of such prestige might come from indirect reciprocity, because donations might be rewarded by outside observers. Indeed, Milinski, Semmann, and Krambeck (2002b) have shown that in an experiment donations to UNICEF are rewarded by other players. This allows for important e¤ects of strategic behavior. If at least some others are indirectly reciprocal, it can pay to strategically build a reputation for being generous. This in turn allows a potential recipient of a donation to increase the revenues by publishing donations. The latter should increase future returns of the donor on his contribution and hence his incentive to give. Apparently charities are aware of the prestige motive (which might be driven by expectations of indirect reciprocity) since it is common practice to announce

[1] We provide a detailed review of the recent literature on indirect reciprocity in Section 4.

[2] In an alternative model of indirect reciprocity, a person that has been helped then helps a third party. Boyd and Richerson (1989) study this form of indirect reciprocity in small cyclical networks.

donors' names and contributions. The interplay of indirect reciprocity and strategic reputation building can thus have substantial impact on economically relevant interaction.[3]

Seinen and Schram (2001) have conducted an experimental helping game (a degenerate game where a donor can help a recipient at a cost smaller than the recipient's bene…t) to explore indirect reciprocity. A subject's previous helping decisions were stored in a so-called image score and the recipient's score was presented to the donor before he decided whether to help or not. This game is nicely suited to study indirect reciprocity because it precludes (in anonymous su¢ciently large groups) any e¤ects of direct reciprocity as opposed to games such as the prisoner's dilemma. Seinen and Schram (2001) …nd that indirect reciprocity is important, which means that many donors base their helping decision on the image score of the recipient. 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.[4]

To assess the interplay of indirect reciprocity and strategic reputation building, it is necessary to clearly distinguish between the possible motives that drive helping choices. An experimental design that achieves this aim thus has to allow us to identify whether observed helping choices are motivated by strategic reputation building or by indirect reciprocity that is not contaminated by incentives for strategic reputation building by the donor, which we call pure indirect reciprocity. In Seinen and Schram observability of the recipient's score implies that the donor's score will be observable in the future as well. Therefore, their design does not allow them to study pure indirect reciprocity uncontaminated by strategic concerns. It is also di¢cult to get a clear idea of the overall impact of strategic reputation building. While they …nd that a substantial part of subjects base their decisions at least in part on their own score, it is not possible to

[3] Andreoni and Petrie (2004) provide further 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. 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 in‡uenced by direct reciprocity as is exempli…ed by e-commerce. Bolton et al. (2004b) and Keser (2002) provide experimental evidence on the importance of reputation mechanisms in environments with repeated one-shot interactions.

[4] Wedekind and Milinski (2000) provide the …rst experimental test of indirect reciprocity, based on only six periods. They …nd 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.

clearly determine the share of helping choices that are due to strategic reputation building.

To disentangle these two e¤ects, we use a helping game where in any period only half of the players have a public image score, the record of their previous behavior. In particular, each subject has a public image score either in the …rst 40 periods of the experiment or in the last 40 periods. This allows us to identify the motivation behind the helping choices. First, since donors without a public image score interact with recipients with a public image score, we can study pure indirect reciprocity uncontaminated by incentives for strategic reputation building. Second, by comparing the behavior of donors with and without public scores, we can evaluate the relative impact of strategic reputation building on the helping rates. Our design allows us to study the latter issue even within subjects.

We test three hypotheses. First, indirect reciprocity is important, i.e. the probability that donors help increases in the recipient's image score. Second, subjects strategically build a reputation, i.e. for any given score of the recipient (including an absent score) the average helping rate of donors with a public image score is higher than that of donors without. Third, strategic reputation building weakens the reciprocal relation, i.e. the dependence of the donor's helping rate on the recipient's score is weaker for donors with a public score than for donors without.

We …nd support for all three hypotheses. There is a clear positive relation between helping rates and recipients' scores for both donors with and without a public score. The latter provides evidence for indirect reciprocity even in the absence of strategic incentives for reputation building, hence for pure indirect reciprocity. Our experiment thus provides, to the best of our knowledge, the …rst evidence of pure indirect reciprocity in the laboratory.[5] The average helping rate of donors with a public score is, however, more than twice the average helping rate of donors without. Hence strategic reputation building plays an important role as well. There is also clear evidence on an individual level: 80% of the subjects help clearly more often when they have a score than when they do not, including 25% who never help when they do not have a public score, but often help when they have a public score. The mode of donors' public scores is at 4 helping decisions out of the last 5 decisions. This score is the expected payo¤ maximizing score in all …ve sessions. In contrast, if scores are not public, the mode is clearly at 0.

[5] We also …nd substantial helping rates (32%) in interactions where neither the donor nor the recipient has a score. This suggests that motives like altruism or e¢ciency concerns play an important role as well.

More generally, our design allows us to draw a clean distinction between non-strategic cooperative behavior and cooperative behavior driven by strategic reputation building. Our results clearly show that both non-strategic and strategic cooperative behavior are of substantial importance.

Besides its importance for understanding the interaction of strategic and reciprocal behavior, 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 (1998a) have conducted simulations of an evolutionary process based on a repeated helping game. They …nd that maximally discriminating players will eventually take over the population.[6] Leimar and Hammerstein (2001), however, show that this result is based on a too restricted initial set of available strategies. They show that 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] Hence, such strategic reputation building could undermine a cooperative society based on indirect reciprocity. The underlying reason is that it is cheaper to keep one's own score at a level that maximizes expected returns by keeping it constantly at the optimal level without paying attention to the recipient's score.

Our experimental results show that about 15% of the population are pure strategist who are not reciprocal. This is the kind of behavior predicted by Leimar and Hammerstein (2001) to invade an indirectly reciprocal population. Furthermore, our experimental results show that payo¤s relative to group averages are clearly higher for strongly strategic subjects (whose helping rates are generally at least doubled when they have a score) than for the weakly strategic subjects (whose helping rates are not substantially higher

[6] For a helping game where the image score is based only on the last decision, Nowak and Sigmund (1998b) have shown analytically that discriminators outperform altruists and defectors but that cycles occur due to invasion of discriminator populations by altruists, which in turn yields an advantage to defectors.

[7] Nowak and Sigmund (1998a) also study strategies that use both the donor's and the recipient's score. For a case with perfect information they …nd that a type who is maximally discriminating with respect to the recipient's score but in addition only helps if his own score is threatening to fall below this threshold, is most frequent. Leimar and Hammerstein (2001), however, argue, using comparable simulations, that this result requires either a substantial in‡uence of genetic drift or a very small cost of helping. Even in the latter case, cooperative image scorers are invaded by sophisticated strategic subjects who build, based on previous experience, a belief whether it pays to have a positive image score.

when they have a score). They are also higher for the non-reciprocal than for the reciprocal subjects (though not signi…cantly). These results are consistent with the invasion argument by Leimar and Hammerstein (2001) and cast some doubts on the evolutionary explanation for indirect reciprocity suggested by Nowak and Sigmund (1998a).

Finally, our experiment also provides a test of di¤erent reciprocity models. While the models by Rabin (1993) and Dufwenberg and Kirchsteiger (2004) cannot be generalized from direct to indirect reciprocity, the models by Levine (1998) and by Charness and Rabin (2002) can also be interpreted as models of both direct and indirect reciprocity. In the model by Levine (1998) the intensity of empathy towards another person depends on the altruism of that person. This altruism can, for example, be inferred from the behavior towards a third person, implying indirect reciprocity. In Charness and Rabin (2002) people are assumed to exhibit "concern withdrawal", i.e. they assign a lower (or even negative) weight to another person's payo¤ in the own utility function if that person has not behaved well, which could, for example, been by the mistreatment of a third person.

The paper proceeds as follows. Section 2 presents the helping game and the experimental design. The results are presented in Section 3, followed by a discussion of related literature 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 (1998a) and Seinen and Schram (2001). There were 16 subjects in each of our …ve 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 of 6 "Points", which yielded a bene…t of 15 Points for the recipient. The recipient had no choice to make.

Each subject had a public score either in the …rst 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 and hence it precluded that donors with and without score behaved di¤erently because they considered themselves advantaged or disadvantaged.[8] A public score consisted of the number of times the subject had helped and had not helped in the last 5 times as a donor. In case the subject had so far been in the role of the donor less than 5 times, the score consisted of the total number of help and not help decisions so far. When the recipient had a public score, the donor was informed about this score before making the decision to help. A subject with a public score was also informed about her or his own score. In case the subject did not have a public score, no score information was displayed (but of course, subjects could easily keep track of their own score). The experimental software did, however, also record the private scores to allow an easy comparison of the choices with and without public score. A 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. It is, however, impossible with our information structure 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.

Of course, classical game theory based on common knowledge of narrow sel…shness and rationality implies no help in the one-shot game and by backward induction for the …nitely repeated game. In an in…nitely repeated helping game, however, it is a Nash-equilibrium when all players choose the strategy to help if and only if they have always been helped and all the recipients that they observed so far had a perfect score. On the equilibrium path all players would then help. This is, however, less straightforward than such trigger strategy equilibria in, e.g., a prisoner's dilemma game. If a player does not help in one period, this is only observed by the current partner, and hence not all players will immediately switch to not helping. As a consequence a player who observed a deviation and who has a public score will in early periods not switch to not helping, because this would immediately a¤ect his own score, while cooperation

[8] While subjects might have considered having a score advantageous or disadvantageous, they knew that they would not be advantaged or disadvantaged over the whole course of the experiment. In a design that simply mixed subjects with and without score, behavioral di¤erences between them might be caused by feelings of being assigned to the inferior role.

will break down in general only after several periods. Hence the break down of cooperation will come with delay. The most important aspect of such an equilibrium is, however, that it requires full helping at all times. If all subjects used the same cut-o¤ strategies which is less demanding than a perfect score, then there are two options: …rst, these strategies could condition only on the current recipient's score. If this is the case, a player would only have an incentive to keep his own score just at this cut-o¤ level, but no incentive to be reciprocal, i.e. to apply a cut-o¤ strategy himself. Second, the strategies might also be sensitive towards whether other subjects are reciprocal. In that case a subject would stop being reciprocal when at least once she is not rewarded although her score is above the cut-o¤ level (for short-term reasons, she might switch temporarily to score-optimizing behavior and later to not helping). This, however, would lead to an eventual complete breakdown of helping if ever a subject is not reciprocal. This in turn implies that there will never be a deviation from always helping, because as long as all have helped, a decision not to help would also mean not being reciprocal, which would in turn lead to a breakdown of reciprocity. The main conclusion is that even in an in…nite game there cannot be an equilibrium where players behave reciprocally but not all players help all the time. Even if one allows for errors non-perfect cooperation cannot be an equilibrium, because if players are tolerant towards other players' errors, this would create an incentive not to help at least occasionally.

Similarly, one can consider a …nitely repeated game with a small number of intrinsically motivated reciprocal players and equilibria where sel…sh players try to mimic such a reciprocal player. This, however, does not work, because the score does not provide information about a player's reciprocal behavior (this would require higher order information), but only about his helping behavior. Hence strategic players will optimize their own score, but not behave reciprocally (unless the intrinsic reciprocal players also react towards whether others treat them reciprocally, as discussed above). If su¢ciently many other players help in order to maximize their own score, it would even pay not to help at all and hence cooperation would break down immediately (unless the number of intrinsically motivated reciprocal players is rather large).

2.2 Experimental Procedures

The experimental software was programmed in z-Tree (Fischbacher, 1999) 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 …elds from the University of Zurich and the ETH Zurich and were recruited by telephone. They were randomly assigned to seats in the laboratory. Instructions were provided in written form and participants could read through them at their own pace (see the appendix 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. More neutral terminology would instead appear plainly arti…cial. At the end of the instructions there were …ve 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 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 payo¤ and total payo¤ 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 score of the recipient or that the recipient did not have a public 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 and without public score were switched and the scores were set to 0.

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, corresponding to a show-up fee of 10 Swiss Francs. 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).[9]

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

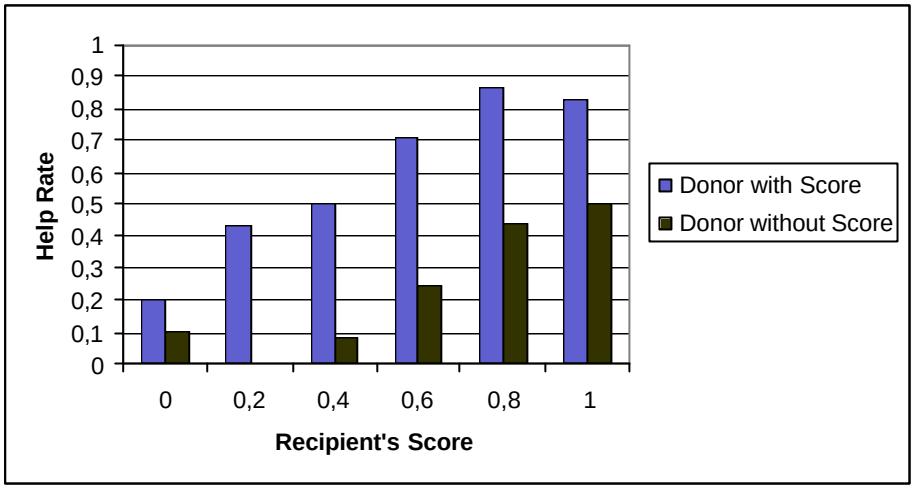


Figure 1: Donors' average help rate for all recipients with a public score based on at least one decision

3 Experimental Results

The overall experimental results are displayed in Figures 1 and 2, which show the average helping behavior of donors with and without score for di¤erent public scores of the recipients. The latter are represented as the relative number of helping choices, i.e. the number of helping choices in the last …ve help decisions divided by …ve, or if the score was based on less than …ve choices, the number of helping choices divided by the total number of choices (and rounded to multiples of 0.2). In Figure 1 all choices where recipients had a public score based on at least one decision are included, in Figure 2 only the decisions where recipients had a full score, i.e. a score based on …ve decisions are included. 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 and only including recipients with a full public or private score). Table 1 shows in particular that helping rates are quite high (32%) even when neither the donor nor the recipient has a score, i.e. in a situation where indirect reciprocity and strategic reputation building cannot play a role. This suggests that additional motives like altruism or e¢ciency concerns are important as well.

Result 1: Donors both with a public and a private score help recipients with a higher score more often. The helping behavior of donors with a private score implies in particular that non-strategic cooperative behavior is important.

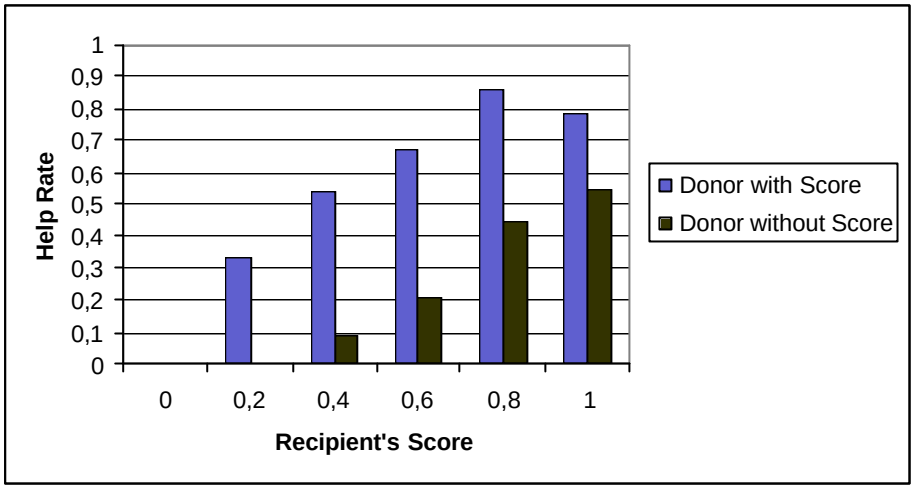


Figure 2: Donors' average help rate for recipients with full public score

Figures 1 and 2 provide immediate support for the …rst hypothesis that donors are indirectly reciprocal and in particular for the existence of pure indirect reciprocity. The helping rate of donors both with and without public score clearly increases with the image score of the recipient, although the relation is monotone only in the case where donors do not have a score and recipients have a full score. A straightforward statistical test con…rms the signi…cance of this positive relation. The H_0 hypothesis is that donors do not condition their helping decision on the recipient's score. We conduct a simple binomial test for this hypothesis based on the …ve sessions as independent observations. To obtain an estimate whether there is a positive relation between the recipient's score and the helping probability we estimate for each session independently a simple linear probability model

$$Pr(Help)_j = const + ®¢ RsScore_j + "_j;$$

where Pr(Help) is the probability that the donor helps the recipient, RsScore is the recipient's score and $"_j$ is a normally distributed error term.[10] Under H_0 in each individual session the probability that the

[10] The linear probability model is clearly not the ideal model to analyze the dependence of the help rate on the recipient's score. Here, however, we only use it to provide us with an input for the simple non-parametric test whether there is a positive relation between the recipient's score and the help probability. For this test it is also irrelevant if coe¢cients are not signi…cantly di¤erent from 0, as long as they are positive.

	Session	1	2	3	4	5	Total
R with	D with score	72%	78%	70%	66%	86%	74%
score	D without score	45%	42%	22%	27%	49%	37%
R without	D with score	72%	66%	63%	66%	75%	69%
score	D without 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

estimated coe¢cient for the recipient's score is positive is $\frac{1}{2}$ (actually slightly smaller, because of a small positive probability for a ‡at relation). Hence, since the sessions are independent, the probability for a positive coe¢cient in all …ve sessions is (slightly smaller than) $\left(\frac{1}{2}\right)^5 = \frac{1}{32} < 5\%$. Since we …nd a positive coe¢cient in all …ve sessions we can thus reject the H_0 that there is no positive relation at the 5% level.[11] This holds independently of whether we include all recipients with a score or only the recipients with a full score and whether we consider donors with or without score (see Table 2).

The a¢rmative result for the …rst hypothesis can also be derived from a panel data analysis (with the sessions as independent units of observations). We use a random-e¤ects probit model.[12] The model is

$$\Pr(\text{Help})_{it} = \Phi(\text{const} + \beta \cdot \text{RsScore}_{it});$$

with, as above, Pr(Help) the probability that the donor helps the recipient, RsScore the recipient's score, and © the normal cumulative distribution function. We again run the regression separately for donors with and without score, to be able to detect whether a reciprocal relation might be restricted to one group of donors. Table 3 lists the estimates for the coe¢cients, the standard errors, z-statistics and p-values. No matter whether we include all recipients with a score or restrict the analysis to those with a full score, both for donors with and without scores the coe¢cient for the recipient's score is positive and highly signi…cant ($p < 0{:}1\%$): Obviously, the same results occur if we run the regression jointly for all donors with and without score.

[11] 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.

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

	Session	1	2	3	4	5
All recipients with score	All donors	:56 (8:1)	:59 (6:4)	:47 (6:1)	:51 (5:2)	:57 (3:9)
	donors with score	:55 (8:7)	:58 (9:2)	:52 (8:3)	:43 (4:4)	:34 (2:2)
	donors without score	:53 (7:0)	:75 (10:2)	:36 (4:7)	:64 (8:9)	:66 (4:5)
recipients with full score	All donors	:65 (9:0)	:74 (8:8)	:51 (5:3)	:46 (3:9)	:72 (5:7)
	donors with score	:60 (7:6)	:62 (8:3)	:44 (3:7)	:30 (1:3)	:32 (1:3)
	donors without score	:69 (10:8)	:93 (13:7)	:49 (6:8)	:69 (10:8)	:87 (7:2)

Table 2: Coe¢cients for recipient's score in linear probability model for helping choice. Adjusted R^2 (in percent) in parantheses.

The high average helping rate of donors without a public score (37% if recipients have a public score, 32% if they do not) is clear evidence of non-strategic cooperative behavior, because these donors do not have an incentive to help to build a reputation that would bene…t them.

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.

Figures 1 and 2 as well as Table 1 provide clear support for the second hypothesis that donors try to strategically build a reputation. The average helping rate of donors with score is higher than of those without for any score of the recipient (including absent score, as can be seen from Table 1).[13] The same holds for each individual session. In case only recipients with full score are considered, there is only one

[13] If the analysis is restricted to recipients with full score, then the helping rate of both donors with and without score is 0 for recipients with a score 0. There are, however, only 13 interactions with a recipient with a full score of 0. 12 of these are with the same subject and hence all in session 1. Furthermore, since the helping rate for donors without score is already 0 for recipients with a score of 0.2, this tie appears to simply result from censoring.

			Coe¢cient	Std. Error	z	Pr > jzj
Donors	All R with	®	1.6028	0.2277	7.04	0.000
with	score	const	-0.4694	0.1758	-2.67	0.008
scores	R with	®	1.517	0.2756	5.50	0.000
	full score	const	-0.4547	0.2090	-2.17	0.030
Donors	All R with	®	1.7893	0.2238	8.00	0.000
without	score	const	-1.6893	0.1854	-9.11	0.000
scores	R with	®	2.328	0.2854	8.16	0.000
	full score	const	-2.1379	0.2661	-8.03	0.000

Table 3: Coe¢cients for the recipient's score (®) in probit model for the help choice

tie, in session 1 for a score of 0. Such a dominance of the helping rate of donors with a score over that of donors without a score occurs in one session under the H_0 hypothesis that strategic reputation building is not relevant with probability less than $\frac{1}{2}$: Thus the fact that it holds in all …ve sessions allows us to reject the H_0 at the 5% level.

This result as well is supported by a panel data analysis. We extend the above model to

$$\Pr(\text{Help})_{it} = \Phi(\text{const} + \beta \cdot \text{RsScore}_{it} + \gamma \cdot \text{DWithScore}_{it});$$

with DWithScore a dummy that takes the value 1 if the donor has a score and 0 otherwise and the other variables as above. Both running the regression for all recipients with a score or only for those with a full score, yields a highly signi…cant ($p < 0{:}1\%$) coe¢cient for the dummy, supporting the second hypothesis (see Table 4).

As can be seen in Table 1, in each session for both recipients with and without score the helping rates of donors with score is about twice the helping rate of donors without score. Hence the impact of strategic reputation building is not only statistically signi…cant, but also of substantial magnitude. On the other hand, both for donors with and without score, the average helping rate is only slightly (about 5%) lower if the recipient does not have a score than when he does. Hence recipients without a score are treated roughly as if they had an average score. This indicates that donors do not create any self-serving beliefs

		Coe¢cient	Std. Error	z	Pr > jzj
All R	®	1.6513	0.1605	10.29	0.000
with	¯	1.1277	0.0720	15.66	0.000
score	const	-1.618	0.1535	-10.54	0.000
R with	®	1.8824	0.1974	9.53	0.000
full	¯	1.0852	0.0812	13.36	0.000
score	const	-1.8302	0.1623	-11.28	0.000

Table 4: Coe¢cients for the recipient's score (®) and for a dummy whether the donor has a score (¯) in probit model for the help choice

that recipients about whom they do not receive information are not helpful, which in turn could justify not helping these recipients.

The importance of strategic reputation building is also very vividly illustrated by Figure 3 which shows the distribution (absolute frequencies on top of bars) of donors' full (public or private) scores.[14] The dark bars show the distribution for donors with private scores and the light bars for donors with public scores. In the former case, the mode is with about 40 % 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 relative score of 0.8 (i.e. 4), with few cases of scores below 0.6 and hardly any below 0.4.[15] Interestingly, in all sessions the score that maximizes expected payo¤s for the observed helping rates is 0.8 (see Table 5). This implies that either many subjects manage to strategically optimize their score, or that many subjects use the most prevalent score as a cut-o¤ point. As Table 5 shows, it appears in general optimal to keep the score at or

[14] We included all full scores following the donor's decision, except for scores resulting from a donor's decision in the last period, because in that case the resulting score could not possibly be relevant for future interaction. The total number of the included scores is 2480, 1227 where the score is public and 1253 where the score is private (the di¤erence 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). Since participants could be a donor (or a recipient) several periods in a row, some of these scores may have never (or several times) been observed.

[15] Of the 19 full scores of 0, 15 come from the same subject, the only pure egoist. In all …ve sessions the mode for private scores is 0. For public scores the mode is 0.8 in three sessions. In one session the mode is 0.6 and in one session it is 1, with 0.8 being the second most frequent score in both cases.

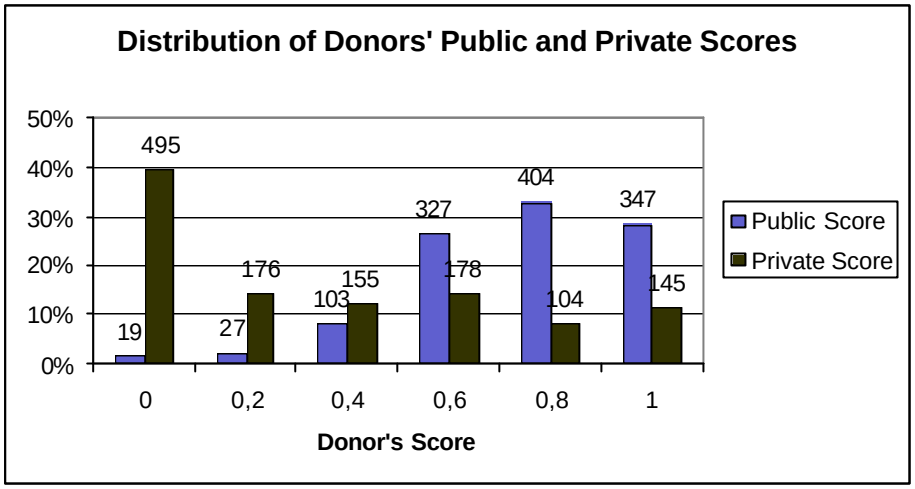


Figure 3: Distribution of public and private (post-decision) donors' scores for all interactions where the donor had a full score (except for following donors' decisions in the last period). Absolute numbers appear on the top of the bars.

close to 0.8 (this analysis does not take time trends into account, the optimal score is calculated based on average help rates over the whole phase with full scores). The high returns on a score of 0.2 in session 2 and on 0.4 in session 5 can be attributed to low numbers of observations (1 and 3, respectively).

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 also study the importance of strategic reputation building on an individual basis by comparing the helping rates with and without score within subjects. Table 6 shows a classi…cation of subjects. We call a subject strategic if the helping rates are generally higher in the part of the experiment where the subject has a score than in the part where he or she does not. A subject is called strongly strategic if the helping rates with score are in most cases at least twice the helping rate without score.[16] Finally, a pure strategist never helps when he or she does not have a score, but does so

[16]For the classi…cation of both strategic and strongly strategic, we allowed deviations from the criteria for one value of the recipient's score and in that case required it to hold strictly for at least two values of the recipient's score. In particular we required the criterium to hold for the case where the recipient did not have a score, because the number of observations was much higher than for any single recipient's score.

Score \ Session	1	2	3	4	5	Total
0	0	0	0	0	0	0
0.2	-0.6	2.9	0.57	-0.6	-0.6	0.57
0.4	0.55	0.91	1.11	1.03	2.3	1.03
0.6	2.1	1.2	1.22	0.87	1.8	1.39
0.8	2.52	3.01	1.51	2.63	2.85	2.42
1	2.07	2.48	1.11	0.81	2.62	1.92

Table 5: Average expected return per period (in Points) from keeping a certain score, based on average help rates over the whole phase with full scores

several times otherwise.[17] There are several special cases of non-strategic subjects. Simple egoists never help, simple altruists always help and negatively reciprocal altruists always help when the recipient does not have a score or when the recipient's score is above some cut-o¤ level, but not for a lower score. A subject is classi…ed as reciprocal when there is a clear positive relation between the recipient's score and the helping rate.[18]

As Table 6 shows, the majority of subjects is clearly strategic, there are only 16 (20%) non-strategic players, but even 20 pure strategists and 23 strong strategists. The distribution of types does not depend substantially on whether the subjects have a score in the …rst or the second half of the experiment. It also does not di¤er a lot between sessions, but the concentration of pure and strong strategist is especially high (75%) in session 3. Interestingly, session 3 also has the highest number of reciprocal subjects (62.5%). There is only one simple egoist, and even more surprisingly only one negatively reciprocal altruist, which

[17] Note that the pure strategists were not included in the strong strategists.

[18] We allowed one exception from the criterium 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 ‡at helping rate in case the donor did not have a score was allowed if the helping rate in case he did have a score showed a clear positive relation. For most subjects, the classi…cation was straightforward, because there was either a clear monotone relation or none at all.

	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 6: Classi…cation 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 …fth rows.

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 0.4. Overall, the number of these simple types is much lower than in Seinen and Schram (2001), who …nd 11% egoists and 36% altruists.[19] Thus we …nd a higher share of subjects who are reciprocal but also a higher share that are strategic. Interestingly, 40% of the pure strategists and 52% of the 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 and hence instead of just exploiting the cooperative system based on indirect reciprocity, they also stabilize it.[20] 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.

Result 4: Strategic incentives to build a reputation weaken the reciprocal behavior

To test our third hypothesis that the reciprocal relation is weaker for donors with score than for donors without, we have to de…ne measures for the reciprocal relation. There are two way to operationalize this

[19] The numbers are, however, not perfectly comparable, because we use a di¤erent classi…cation than Seinen and Schram (2001). In particular, our design allows us to more clearly detect strategic players and hence our classi…cation is …ner in that respect.

[20] This can be seen as being strategic on a higher level, because due to the matching procedure, donors could pro…t from inducing others to help, either by later being matched with them again or by indirect e¤ects.

concept. First, since donors with public scores have to be concerned about their own score, they cannot solely condition their helping decision on the recipient's score. Therefore, the correlation between their helping decision and the recipient's score should be weaker. A second measure, that we use here, is the slope of a linear regression between helping decision and the recipient's score. This slope is a measure of the reciprocity of the donor. Consequently, it expresses the average incentive for the subjects with public score. It is easier for the subjects with private scores to establish this incentive (because subjects with public score might also help recipients with low score in order to keep up their own score). Therefore, we expect them to have a steeper slope. Indeed, if we focus on the recipients with a full score (which appears appropriate, because the preceding interactions are subject to more random influence due to initial experimentation of donors), then Table 2 shows that in each session the coefficient of the recipient's score in the linear probability model is higher for donors without score than for donors with score. Hence the relation between the recipient's score and the helping probability is stronger for donors without score than for donors with score in each session, and thus overall is significantly stronger for the donors without score at $p = 5\%$: We can also address this question by a second measure, the adjusted R^2 for each regression. As is shown in Table 2, the adjusted R^2 is higher for the donors without a score than for the donors with a score in each session, and hence overall, the recipient's score is a (significantly) better predictor for the helping probability if the donor does not have a score.[21] If we consider all recipients with a score, however, both these tests fail and hence the overall support for the third hypothesis is weaker than for the first two hypotheses.

These result are again supported by a panel data analysis. We further extend the above model to

$$\Pr(\text{Help})_{it} = \Phi(\text{const} + \beta \cdot \text{RsScore}_{it} + \gamma \cdot \text{DWithScore}_{it} + \delta \cdot (\text{DWithScore} \times \text{RsScore})_{it})$$

where the interaction term DWithScore × RsScore captures the additional effect of the recipient's score in case the donor has a score and the other variables are defined as above. As shown in Table 7, the coefficient for the interaction term is significantly ($p < 5\%$) negative if we restrict the analysis to the recipients with

[21] The fit of the model is in general not very good because donors' behavior is quite heterogenous. This, however, rather strengthens the result since still a clear effect can be detected.

		Coe¢cient	Std. Error	z	Pr > jzj
All	®	1.7467	0.2248	7.77	0.000
R	¯	1.3043	0.2547	5.12	0.000
with	°	-0.2275	0.3192	-0.71	0.476
score	const	-1.7648	0.1858	-9.50	0.000
R	®	2.3425	0.2832	8.27	0.000
with	¯	1.7892	0.3102	5.77	0.000
full	°	-0.9319	0.3945	-2.36	0.018
score	const	-2.1936	0.2294	-9.56	0.000

Table 7: Coe¢cients for the recipient's score (®), for a dummy whether the donor has a score (¯) and for the interaction term (°) in probit model for the help choice

full score, supporting the third hypothesis that the impact of the recipient's score is weaker if the donor has a score. If we consider all recipients with a score, however, the coe¢cient is again negative, but far from signi…cant ($p > 40\%$):

We …nd furter support for stratgic reputation building by extending our probit model to

$$\Pr(\text{Help})_{it} = \Phi(\text{const} + \Re \cdot \text{RsScore}_{it} + \bar{} \cdot \text{DWithScore}_{it} + \circ \cdot (\text{DWithScore} \pounds \text{RsScore})_{it} + \pm \cdot \text{DsScore}_{it} + \acute{} \cdot (\text{DWithScore} \pounds \text{DsScore})_{it})$$

including the donor's own score DsScore and an interaction term DWithScore £ DsScore. The …rst term essentially captures individividual di¤erences in the propensity to help. Donors who help more often have a higher score and hence in general the current score and the probability to help should be positively correlated. The interaction term then captures to what extent this relation is weakened if the donor has a public score. The results are presented in Table 8 for the case of donors and recipients with full score.[22] We …nd again all the e¤ects from above. In addition, the positive e¤ect of the donor's own score is consistent with individual di¤erences in the propensity to help, because this implies that some donors

[22] The results do not change substantially if we include in the analysis all donors with a score and all recipients with a score instead of only those with a full score. The only e¤ect is that the interaction e¤ect between the recipient's score and the dummy whether the donor has a score is weaker, though still signi…cant.

	Coe¢cient	Std. Error	z	Pr > jzj
®	3.3968	0.3807	8.92	0.000
¯	2.5351	0.4511	5.62	0.000
°	-1.7843	0.4786	-3.73	0.000
±	3.0696	0.2298	13.36	0.000
´	-1.6749	0.3541	-4.73	0.000
const	-4.0605	0.3444	-11.79	0.000

Table 8: Coe¢cients for the recipient's score (®), for a dummy whether the donor has a score (¯), for the interaction term (°), for the donors score (±), and for the interaction term (´) in probit model for the help choice

have consistently a higher score and help more often. The interaction e¤ect between the donor's score and the dummy indicating whether he has a public score, however, is signi…cantly 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 in the sense that donors with a low score try to increase their score, but those with a very high score do not mind if it falls a bit.[23] This behavior is rational because the payo¤ maximizing score is at a high, but not the maximal possible level.

Our a¢rmative result on the third hypothesis has a potentially important, though at the current state somewhat speculative, implication. If the share of subjects with public score was higher, but the helping rates (conditional on the recipient's score) for donors with and without public score were not a¤ected, average helping rates would be higher, but the incentives to help in order to build a reputation would be weaker.

One could 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 …nd 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

[23] If we restrict the analysis to donors with a public score, the coe¢cient of the donor's score is still positive and signi…cant, but lower than for donors with a private score.

one observes a low score (though this should strictly be so only in early periods or if subjects are highly myopic, because otherwise the total information one has gathered so far should dominate this period's recipient's score). This potential interpretation, however, appears to be valid only for donors with score, because donors without a score do not have an incentive to …nd out what constitutes a successful score. The support for our third hypothesis, that the relation between recipient's score and helping probability is stronger for donors without a score, thus contradicts this interpretation and suggests that what we see is indeed indirect reciprocity.

Result 5: Strategic subjects obtain signi…cantly higher payo¤s than non-strategic subjects. Reciprocal subjects obtain lower payo¤s than non-reciprocal subjects.

Con…rming straightforward intuition, strategic reputation building pays, whereas reciprocity does not. Table 9 shows the average payo¤s (relative to the average payo¤s in the session) of subjects by being reciprocal and strategic, where we summarized the subjects who were classi…ed as pure or strong strategists as strongly strategic and those that were classi…ed as weakly or non-strategic as weakly strategic. Clearly, the strongly strategic outperform the weakly strategic, 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.[24] More importantly, it pays not to be reciprocal, apparently because being reciprocal distracts from perfectly …ne-tuning one's own score (or, in case of private scores, is a pure waste).[25] This indicates that in an evolutionary game based

[24] If we study the data in a more disaggregated way, we …nd that the payo¤ for the purely strategic is slightly higher than that for the strongly strategic and the payo¤ 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 di¤erence is between strongly strategic and weakly strategic we aggregated the data in two categories for the present analysis.

[25] We see no reason why the individual subjects' payo¤s relative to the session average should be dependent within one session, hence we use the relative payo¤s of individuals as independent observations in the following tests. According to both Mann-Whitney and two-sample t-tests, the di¤erences in relative payo¤s between strongly strategic and weakly strategic subjects are signi…cant ($p < 2\%$ for the non-reciprocal, $p < 1\%$ for the reciprocal and $p < 0{:}1\%$ for the whole sample), but those between reciprocal and non-reciprocal subjects are not ($p > 10\%$ for the strongly strategic, the weakly strategic and the whole sample.) For a stricter test, which does not treat the relative payo¤s of individuals as independent, note that the average relative payo¤s are larger for the strongly strategic than for the 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

	Strongly Strategic	Weakly 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 9: Payo¤s relative to average session payo¤ for pure or strongly strategic versus weakly or non-strategic and for reciprocal versus non-reciprocal players, number of players in the respective category in parantheses.

on this repeated helping game and with the experimentally observed player types, the purely strategic non-reciprocal types would drive out the other types and would eventually undermine the cooperation. Given that the relative payo¤ of the non-reciprocal strongly strategic players is almost twice that of the reciprocal weakly strategic players, the evolutionary process would be quite fast for any su¢ciently payo¤-sensitive dynamic.

Results 6: End-game e¤ects are consistent with the major patterns of behavior.

Figure 4 shows the development of the average helping rates in the …rst 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 value of 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 e¤ect is remarkably consistent with our classi…cation of subjects into purely strategic, strongly strategic, weakly strategic and non-strategic based on the comparison of behavior with and without public score. 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 he or she has a public score. What we observe when we compare 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 classi…ed 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

do better at p = 5%: The non-reciprocal do better than the reciprocal in only four sessions and hence this test also misses statistical signi…cance.

it.[26] The corresponding numbers for the 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 classi…cation of subjects in terms of strategic behavior. Subjects classi…ed as purely or strongly strategic exhibit a clear drop in helping behavior towards the end while those classi…ed as weakly or non-strategic do not.[27]

Furthermore, the end-game e¤ect is almost exclusively restricted to the subjects with a public score, 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. For players with a private score, the helping rate in the last two periods of the …rst 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 …rst 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 and above 60% in all but one previous periods; Second phase: period 39: 41%, period 40: 29%, overall average: 72%, the helping rate is above 45% in all previous periods and above 58% in all but one previous period). 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 di¤erence in their behavior is driven by strategic reputation building that cannot matter in the last period.

Overall, the helping rate is somewhat lower in the second half of the experiment (51%) than in the …rst (56%). In the individual sessions, this e¤ect is not consistent. The di¤erence is reversed in one session and virtually zero (0.3%) in another.

[26] 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.

[27] Overall helping rates do not di¤er dramatically between the di¤erent categories in the phase when subjects have a score. For purely strategic subjects it is 67%, for strongly strategic 68%, for weakly strategic 84%, and for non-strategic 72%. Therefore, the observation concerning end-game e¤ects is not an artifact of di¤erences in overall helping rates.

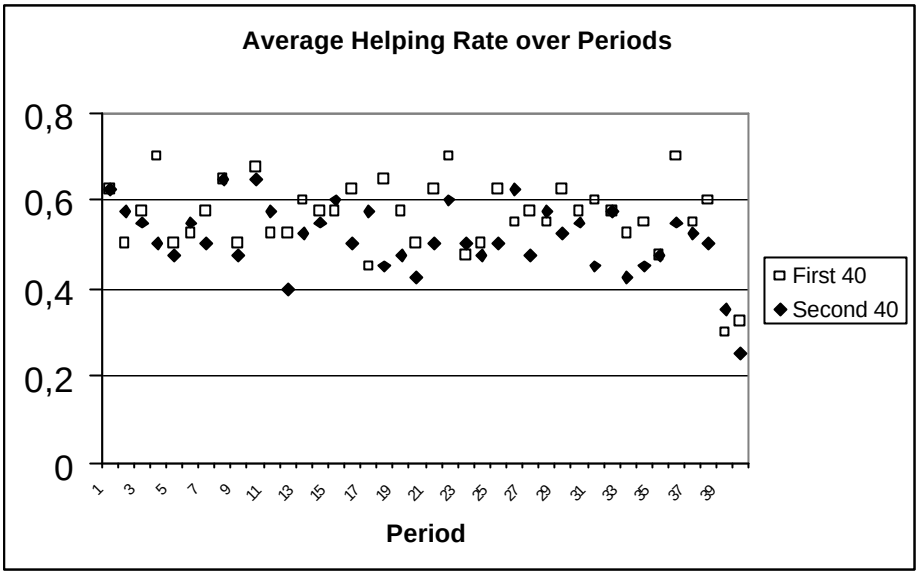


Figure 4: Average helping rates over all sessions for the respective 40 periods of the …rst and the second half of the experiment

Seinen and Schram (2001) analyze the development of di¤erent group norms, which are apparent minimal scores that donors demand from a recipient in order to help him or her and which they also try to achieve themselves. They …nd that over time di¤erent groups tend to adopt di¤erent norms, since di¤erent groups tend towards di¤erent helping rates and di¤erences between groups increase over time. We see relatively little evidence for similar tendencies in our data. In both phases of the experiment, helping rates in the …rst 10 periods di¤er about as much as in periods 21-30. And they even di¤er less in the last 10 periods with the exception of a stronger end-game e¤ect in session 3. That e¤ect can be attributed to the high number of strongly strategic players in session 3 and thus it does not indicate the development of a di¤erent norm, but rather more violations of the norm.

We also …nd some support for the alternative notion of indirect reciprocity that players help who have recently been helped. When comparing the average help rates of donors who have been recipients in the last interaction and have been helped with those who have been recipients in the last interaction and have not been helped (see Table 10), we see that for all possible donor's scores the help rate is higher for the former (one has to consider donors with di¤erent scores separately, because otherwise one might incorrectly

donor's score	0	0.2	0.4	0.6	0.8	1
public score	-	1	0.82	0.91	0.78	0.92
	0	0.64	0.57	0.73	0.55	0.69
private score	0.09	0.32	0.61	0.55	0.78	0.97
	0.03	0.15	0.34	0.44	0.45	0.85

Table 10: Average helping rates for donors who have been recipients in the last interaction and have been helped (top) and not been helped (bottom), by donor's score and score status.

conclude that being helped increases the probability of helping if in fact donors who have helped more frequently in the past do so in the present but are also helped more frequently). For donors with a (high) public score, it is not entirely clear whether this result really indicates such a form of indirect reciprocity or whether it rather results from subjects realizing that a high score has bene…tted them. For donors without public score the e¤ect is also present which provides more conclusive evidence for this form of indirect reciprocity. It is, however, not very consistent in the individual sessions.

4 Related Literature

4.1 Standing Strategy as an Alternative to Image Scoring

Our experiment has pointed out that subjects react to the strategic incentives that are caused by future donors' indirect reciprocity. This implies that they face a trade-o¤ when they meet a recipient who has a low score. If they are motivated by indirect reciprocity they want to punish this player. Punishing by not giving, however, deteriorates the donor's own score. Hence the donor's indirectly reciprocal motivation may be in con‡ict with the strategic incentives implied by future donors' indirectly reciprocal actions. Furthermore, as pointed out by Leimar and Hammerstein (2001), a system based on image scoring is susceptible to exploitation by strategic score optimizers. These problems can only be avoided if higher order information is provided, for instance, information about the score of the previous partners of the current recipient.

Sugden (1986) analyzes a model with higher order information and what he calls "standing strategy". A

player is in "good standing" as long as he has helped or has refused to help a recipient not in good standing, but loses good standing if he does not help a recipient in good standing. Leimar and Hammerstein (2001) show that the standing strategy is evolutionarily stable if the horizon is su¢ciently long and the probability of execution errors is small, but considerably larger than the probability of perception errors. They show in simulations that even in a case where these conditions are not all ful…lled, the standing strategy dominates the population and hence while it is not evolutionary stable, it is more robust than an indirectly reciprocal strategy based on a cut-o¤ strategy.

Bolton, Katok, and Ockenfels (2004a) conducted an experimental helping game where donors in one treatment were informed only about the current recipient's last choice in the donor role, but in another treatment they were also informed about the recipient's score in that interaction. The underlying model of Bolton et al. (2004a) hence corresponds in the …rst treatment to image scoring models, while in the second treatment it is conceptually closer to that analyzed by Sugden (1986). A crucial di¤erence, however, is that in the situation analyzed by Sugden, there is an incentive both to punish a defector and to help a player who has punished whereas in the underlying model of Bolton et al. there is no incentive to help a punishing player because if one does not help a recipient who has not helped lately, future partners cannot detect whether one has refrained to help an exploiter or a punisher. Hence cooperation based on good standing is more robust than that based on the second order information in the underlying model of Bolton et al.

In Bolton et al. (2004a) the presence of second order information clearly increases cooperation, implying that "deeper" information (looking back deeper into the history of play) fosters helping, whereas our results suggest that "broader" information (a higher share of players with a public score) might also have adverse e¤ects. The …rst order information has the strongest in‡uence when the second order information was "help", i.e. for the probability of being helped it is most important how one last treated another player who has helped. This would be consistent with the conclusion that donors try to follow a standing strategy. If the current recipient's last recipient had helped the last time he was the donor (and hence was unambiguously in good standing), the current recipient's standing can be determined and thus guide the current donor's decision. In contrast, if the current recipient's last recipient had not helped the last time he was the donor, the information is not su¢cient to determine this player's standing and hence also the current recipient's

standing. Therefore, the information on the current recipient is ambiguous and is thus not expected to have a clear impact on the current donor's decision.

The results of Bolton et al. (2004a) agree with ours to the extent that they …nd that strategic incentives (e.g. costs) matter. In their study this is also shown by a clear end-game e¤ect, similar to our results. Given that in our study the score is based on several periods, one might, however, expect helping rates to drop substantially before the second to last period, which they do not. Bolton et al. (2004a) also …nd that being helped increases the probability of helping in the next period, consistent with the alternative notion of indirect reciprocity suggested by Boyd and Richerson (1989).

Similar to the approach by Bolton et al. (2004a), Milinski et al. (2001) compare treatments with …rst and second order information, but they provide information about all the preceding choices of the recipient and his or her previous recipients, not only about the respective last choices. Furthermore, subjects have the opportunity to track also higher order information, because all decisions are made public and subjects are assigned names throughout the experiment. While Milinski et al. (2001) …nd that second order information is taken into account to some extent, they interpret their data to be overall rather in line with an image scoring strategy than with a standing strategy because observed helping rates (for a speci…c group of players, i.e. those that did not help a recipient who had not helped before and were hence not to blame according to a standing strategy) are not signi…cantly di¤erent from expected helping rates under an image scoring strategy (though they were higher) but are signi…cantly lower than expected helping rates under a standing strategy. It appears that subjects took the information that was explicitly provided (…rst and second order) into account, but were not capable of (or not interested in) keeping track of higher order information.

To us, an e¤ective test of a standing strategy appears di¢cult to design. An experimental design that provides su¢cient information for a standing strategy would either overwhelm subjects with information (if it provides them with the necessary third, fourth and higher order information, as in Milinski et al., 2001) or it would directly suggest to use a standing strategy by providing information about the standing explicitly. Outside the laboratory, people might be able to develop e¢cient mechanisms to keep track of others' standing, whereas in an experiment, time might be too short to develop such an e¢cient mechanism.

In particular, in non-anonymous interaction outside the laboratory, information about another's past might be retrieved from memory via emotional reactions, which appears to be much harder in an experiment where identi…cation is only via …ctitious names or player numbers.

4.2 E¤ects of Reputation across Games

Milinski, Semmann, and Krambeck (2002a, 2002b) study how an image score acquired in one game (or individual decision) in‡uences the behavior in another. In the experiment by Milinski et al. (2002a) a public goods game is combined with a helping game, where donors are also informed about the recipient's choice in preceding rounds of the public goods game. They …nd that contributions in the public goods game deteriorate if …rstly eight periods of the public goods game are conducted and then eight rounds of the helping game. In contrast, if rounds of the public goods game and the helping game alternate, contributions in the public goods game stay at a substantially and signi…cantly higher level. Furthermore, contributions in subsequent rounds of the public goods game decline quickly if subjects are informed that no further rounds of the helping game follow, but do not if subjects are not given this information. These results con…rm that subjects clearly follow the strategic incentives to build a reputation. In the design, however, elements of direct and indirect reciprocity are combined. Rewards and punishment in the helping game of previous behavior in the public goods game are clearly a form of direct reciprocity because the donor was a¤ected by the recipient's behavior in the public goods game. Hence the main result shows that subjects strategically react to the possibility of direct reciprocity, not indirect reciprocity. The experiment provides again, however, clear evidence for indirect reciprocity (and even of a similar impact than that of direct reciprocity), because the decisions in the helping game in‡uence the probability to be helped as much as the decisions in the public goods game.

The experiment by Milinski et al. (2002b) shows that donations made to charity (UNICEF) signi…cantly increase the probability to be helped in a helping game. Hence indirect reciprocity does not only occur within closed groups, but also helping outsiders can improve one's chances of receiving help.

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 non-strategic indirect reciprocity and the impact of strategic incentives. The interaction of donors with and without public score is the fundamental di¤erence to the helping experiment by Seinen and Schram (2001). 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 result purely from a motivation for indirect reciprocity and helping choices that are driven by attempts to improve one's own score.

From a general perspective, our separation between subjects that can strategically build a reputation and those that do not, provides a clean separation between non-sel…sh cooperative behavior (helping by donors with private scores) and strategic cooperative behavior (the di¤erence 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 speci…c perspective, we are the …rst to …nd clear evidence for indirect reciprocity even in the absence of strategic incentives for reputation building, but we also …nd very strong e¤ects of strategic reputation building. Speci…cally, 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.[28] This is in contrast to the indirect evidence by Seinen and Schram (2001) who classify more than half of the players as non-strategic and in particular …nd substantially more simple altruists and simple egoists, types that are virtually absent in our experiment.

The pure indirect reciprocity that we …nd is inconsistent with the approaches towards reciprocity by Rabin (1993) and by Dufwenberg and Kirchsteiger (2004) that cannot be generalized from direct to indirect

[28] An alternative explanation for the observed higher helping rates of donors with a public score could be that some donors want to set an example for what they believe to be the right thing to do. Without a public score only the recipient becomes aware of their choice, whereas with a public score their future donors learn their past behavior, which enhances the scope for setting an example.

reciprocity. It is, in contrast, consistent with the approaches by Levine (1998) and Charness and Rabin (2002) that can easily be reinterpreted as general models incorporating both direct and indirect reciprocity.

We also …nd support for our hypothesis that the subjects who have a public score and hence a strategic incentive to help, exhibit somewhat weaker reciprocal behavior than subjects who do not have a public score. Hence while strategic incentives to build up an image score increase an individual's general propensity to help, it weakens the in‡uence of indirect reciprocity. A potential implication is that in a population with a higher share of subjects with a public score, the incentives for strategic reputation building could be weakened.

Concerning the empirical relevance of the invasion predicted by Leimar and Hammerstein (2001), we clearly …nd strategic non-reciprocal players who also receive higher payo¤s than other types. This casts some doubts on the evolutionary explanation for cooperation based on indirect reciprocity suggested by Nowak and Sigmund (1998a) because the types predicted to undermine the cooperation by exploiting the system are clearly present and more successful. Put di¤erently, 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 are out there, so any simulation or evolutionary model that tries to explain some phenomenon in the human society 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 without a public score contradicts the evolutionary model by Nowak and Sigmund. In their model indirect reciprocity evolves where players can build a reputation. The donors without public 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.

How can we explain the coexistence of pure indirect reciprocity and strategic reputation building in the laboratory? It might be that indeed life outside the laboratory often does provide higher order information

that allows people to apply, for example, a standing strategy. In such a rich environment, it is possible to reward other people who not only have helped often, but who have been reciprocal and not to help people who help indiscriminately and hence appear helpful but not reciprocal. In such an environment, a strategic player would have to behave reciprocally. If intrinsically motivated reciprocal people and strategic players interact in such an environment, they would not only have similar image scores, as in our experiment, but would be indistinguishable in behavior and hence in payo¤s so that strategic players do not have an evolutionary advantage. In our laboratory situation intrinsically motivated reciprocal players might try to base their reciprocal behavior on the best information they can get, which is the image score. In contrast, strategic players realize the strategic incentives to optimize their image score.[29][30] If the world had an information structure like our experiment and behavior were driven by evolution, indirect reciprocity would not survive. The implication is that our experiment, and hence the model by Nowak and Sigmund does not correctly re‡ect the information structure of the world outside the lab and hence that an evolutionary explanation of the coexistence of indirect reciprocity and strategic behavior requires models with a richer information structure. Alternatively, one might argue that the exploiters are a relatively new phenomenon and that we are still in the take-over process. While this view might be correct, we consider it too depressing to share.

As a …nal result, 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 …ttest 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.

[29] This is obviously not a complete explanation, because some players are clearly strategic, but still help (though at a lower rate) when they do not have a public score.

[30] The reciprocal players could prevent being exploited by strategic players, but only at a high cost. If they play a trigger strategy and stop being reciprocal once they have not been treated reciprocally (have not been helped although their score is above the threshold), strategic players would have an incentive to be reciprocal. But in this case, a single deviation would lead to a complete breakdown of cooperation.

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 Con…dentiality: 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] Blount, Sally, 1995. "When Social Outcomes Aren't Fair: The E¤ect of Causal Attributions on Preferences." Organizational Behavior and Human Decision Processe 63, 131–144.

[5] Bolton, Gary E., Katok, Elena, and Ockenfels, Axel, 2004a. "Cooperation among Strangers with Limited Information about Reputation." Journal of Public Economics, forthcoming.

[6] Bolton, Gary E., Katok, Elena, and Ockenfels, Axel, 2004b. "How E¤ective Are Online Reputation Mechanisms? – An Experimental Study." Management Science, forthcoming.

[7] Boyd, R. and Richerson, P. J., 1989. "The Evolution of Indirect Reciprocity." Social Networks 11, 213–236.

[8] Charness, Gary, 2002. "Attribution and Reciprocity in an Experimental Labor Market." Forthcoming in Journal of Labor Economics.

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

[10] Cox, James, 2004. "How to Identify Trust and Reciprocity." Games and Economic Behavior, 46, 260–281.

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

[12] Falk, Armin and Fischbacher, Urs, 1999. "A theory of Reciprocity." Working paper No. 6, Institute for Empirical Research in Economics, University of Zurich.

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

[14] Fischbacher, Urs, 1999. "Z-Tree: Zurich Toolbox for Readymade Economic Experiments." Working paper No. 21, Institute for Empirical Research in Economics, University of Zurich.

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

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

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

[18] Levine, David K., 1998 "Modeling Altruism and Spitefulness in Experiments." Review of Economic Dynamics, 1(3), 593–622.

[19] Milinski, Manfred, Semmann, Dirk, and Krambeck, Hans-Jürgen, 2002a. "Reputation Helps Solve the 'Tragedy of the Commons'." Nature 415, 424–426.

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

[21] Milinski, Manfred, Semmann, Dirk, Bakker, Theo C. M., and Krambeck, Hans-Jürgen, 2001. "Cooperation through Indirect Reciprocity: Image Scoring or Standing Strategy?" Proceedings Royal Society London: Biological Sciences 268, 2495–2501.

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

[23] Nowak, Martin A. and Sigmund, Karl, 1998b. "The Dynamics of Indirect Reciprocity.". Journal of Theoretical Biology 194, 561–574.

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

[25] Seinen, Ingrid and Schram, Arthur, 2001. "Social Status and Group Norms: Indirect Reciprocity in a Helping Experiment." Working paper, CREED, University of Amsterdam.

[26] Sugden, R., 1986. The Economics of Rights, Co-operation and Welfare. Oxford: Basil Blackwell.

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

Appendix: Instructions (Not intended for publication)

(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

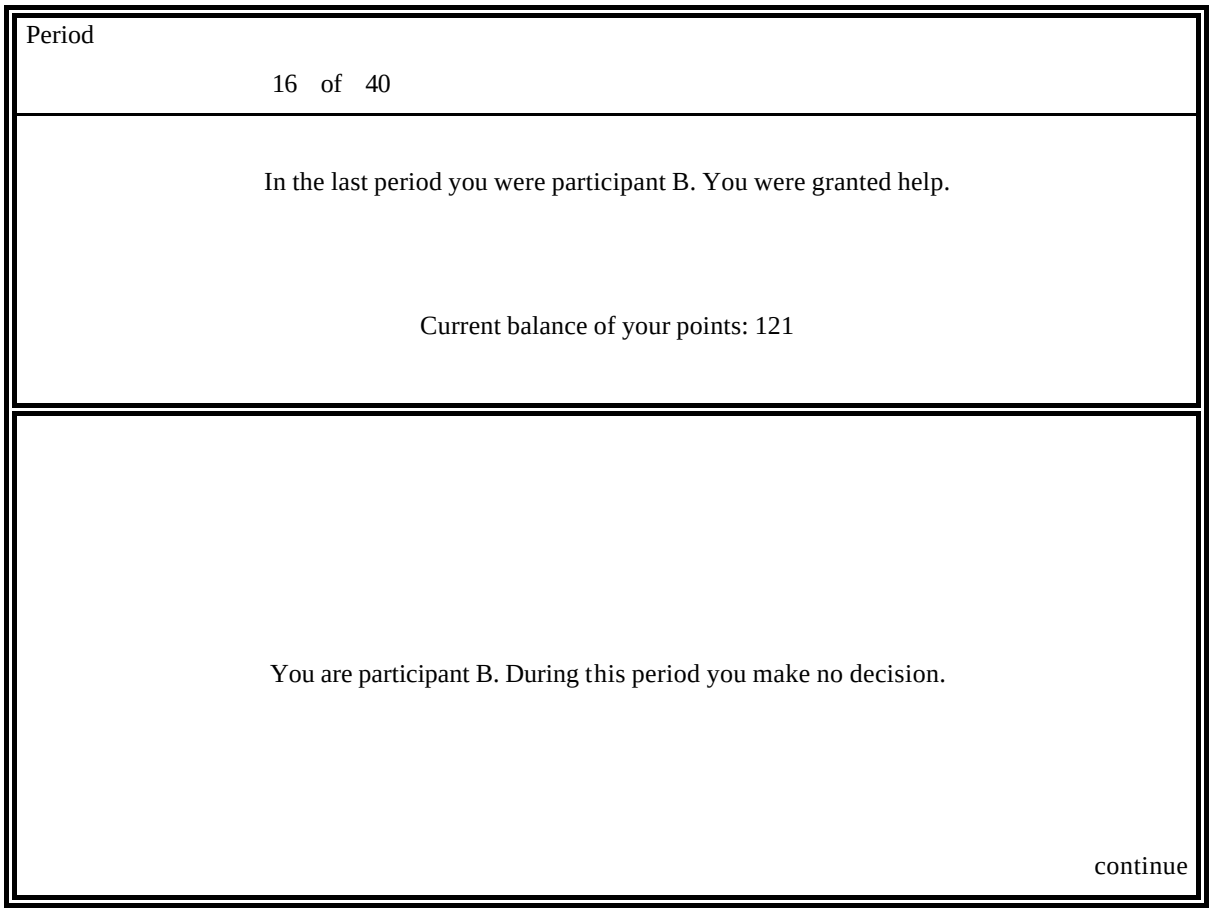
Your decisions will be revealed to your future participants A

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



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?