Keeping up appearances: Experiments on cooperation in social dilemmas

van den Broek, E.M.F.

Publication date
2014

Chapter 2: Direct and indirect reciprocity

2.1 Introduction[3]

Humans are allegedly the 'champions of reciprocity' (Nowak and Sigmund, 2005). Not only do we often return favors, we even 'help' (i.e., direct kind but costly acts towards) people who have been kind to others (Alexander, 1987). These two reciprocal strategies, which have been called direct and indirect reciprocity, respectively, ensure the survival of cooperative individuals by channeling help towards them. Direct reciprocity has been shown to be a stable strategy in small groups of individuals interacting repeatedly (Trivers 1971; Binmore, 1992). In sufficiently large groups, however, repeated interaction may be rare and the probability that two members meet again can be very low. Therefore studies on indirect reciprocity focus on groups in which the possibility of direct experience with a partner's previous choices is negligible (Rosenthal, 1979). Presumably, however, intermediate sized groups were most common in our prehistory (Kelly, 1995). In this case, an individual (A) considering whether or not to help another individual (B) may resort to two kinds of information about B's previous actions. In principle there is information available on how often B has helped A in previous interactions (information for Direct Reciprocity, *IDR*) and information on how often the recipient has helped third parties (C) (information for Indirect Reciprocity, *IIR*). In practice, information about behavior towards third parties is likely to be incomplete or noisy, however, either because of intentional spreading of false information (Hess and Hagen, 2006) or because accuracy simply may decrease through miscommunication. People may therefore treat the two sources of information differently[4].

One way to quantify an individual's willingness to cooperate is by means of a numerical proxy for reputation, i.e. an image score, as explained in Chapter 1. Cooperative strategies discriminatively responding to such a score can arise and are conserved in an evolutionary

[3] Parts of this chapter have been published as "Molleman L., Van den Broek E., Egas M. (2013) Personal experience and reputation interact in human decisions to help reciprocally".

[4] One study shows, counter-intuitively, that gossip influences people's behaviour more than factual information (Sommerfeld et al., 2007).

setting (Nowak and Sigmund, 1998). The approach prompted many follow-ups, theoretically as well as experimentally (for a summary see Nowak and Sigmund, 2005). The most rigorous tests of image scoring show that people indeed base their decision to help on their partner's score (Seinen and Schram, 2006; see also Wedekind and Milinski, 2000). However, agents optimizing their own score pose a challenge to scoring models (Leimar and Hammerstein (2001). This proved to be more than just a theoretical problem, as shown by Engelmann and Fischbacher (2009). In their experimental study, agents had a publicly observable score only in the first or last half of the experiment, which precludes strategic reputation formation in the other half. Non-reciprocal, strategic reputation builders (those that based their choices mainly on their own reputation) earned on average more money than reciprocal players who cooperated based on the other's score. Still, 32 % of cooperative acts was purely reciprocal and uncontaminated by strategic considerations for the own score. In sum, information about previous' behavior of partners increases the likelihood that help is given, but the extent of the impact of this information depends on the strategic incentives for giving it and on the costs of giving (Bolton et al., 2005).

To date only a limited number of studies has investigated the possibility that behavior in an indirect setting might play a role in a subsequent direct interaction, or vice versa, as proposed by Panchanathan and Boyd (2004). In an alternating public goods game (PG) combined with an indirect reciprocity game, contributions in the PG remain high with repetition if they are subsequently disclosed to partners in the indirect reciprocity game (Milinski et al., 2002). In a similar setting, disclosure of PG contributions amplifies generosity in a prisoners´ dilemma scenario (Wedekind and Braithwaite, 2002). These findings show that reputation is an effective instrument to enhance cooperative behavior and even transfers from one environment to another.

It is difficult to conclude from the evidence above whether or not people distinguish between direct and indirect information about others' past choices. To the best of our knowledge there are only two studies that try to tease apart the effects of IDR and IIR on helping, but the results are inconclusive. For example, Dufwenberg et al. (2001) present a trust game experiment where a receiver can reward either the sender, or, as a treatment variable, some other donor.

The latter treatment induces lower investments and, surprisingly, higher repayments. Both effects are statistically insignificant, however. This finding shows that, if anything, indirect reciprocity induces higher return rates. This counterintuitive finding is corroborated by a second study, a one-shot sequential gift exchange game in which direct (if A helps B, B helps A), indirect (if A helps B, C helps A) and generalized (if A helps B, B helps C) reciprocity are compared. Return rates are significantly higher in the generalized reciprocity treatment than in the indirect or even the direct reciprocity treatment (Stanca, 2009). This shows that if people are confronted with a situation in which they can only direct help towards a third party, they do so. As to the comparison between IDR and IIR, in the abovementioned studies people have been observed to use direct and indirect reciprocity interchangeably, directing more help towards helpful individuals in general.

In this chapter, we address a related but different question. We consider a situation where B has to decide whether or not to help A. A has interacted with B in the past, so B has her own experience, but A has also (far more often) interacted with various C's. We are interested in how B's distinguish between their own direct information about A and the indirect information they may have about A's behavior towards third parties.

Theoretical work on the evolutionary success of either form of reciprocity yields various conditions depending on the cost-benefit ratio of helping. A necessary condition that has been derived for direct reciprocity to evolve is that the probability of meeting the partner again in the future must be larger than the cost/benefit ratio (Axelrod and Hamilton, 1981). This is simply a special case of the condition for the evolution of indirect reciprocity, which reads that the probability of knowing the other's reputation must be larger than the cost/benefit ratio (Nowak and Sigmund, 1998). In situations with intermediately large groups, indirectly reciprocal strategies (that base a choice on third-party information) and directly reciprocal strategies (that base decisions on own experiences) can co-exist. Simulations have shown that as the number of encounters between two specific agents increases, for instance because of a smaller group size (in other words, as the probability of meeting a partner again is high relative to the probability that the partner knows your image score), strategies based on direct reciprocity are employed by a larger part of the population (Roberts, 2008). This result

suggests that selection favors strategies using direct information (since they have on average a higher fitness) as the relative reliability of indirect information decreases. The above simulation study disregards strategies that employ both types of information, however.

This prompts three behavioral questions:

(i) Given that the two types of information are simultaneously available, which information do people rely on?

(ii) Do people react differently to the two types of information?

(iii) Do people display preferences for one kind of information that translate into differences in earnings?

The present study seeks to answer these questions in environments with reliable and unreliable third party information. Given the experimental literature mentioned above and Roberts' (2008) simulation results about a gradual shift in relative strength of different strategies in relatively large groups, we conjecture that evolution will favor strategies that take both types of information into account. This hypothesis is first tested by means of an evolutionary simulation that shows that strategies combining direct and indirect reciprocity can invade Robert's disparate strategies. They do not always displace them completely, as they may co-exist with one-dimensional strategies. We find that no particular combination of strategies is favored by selection. Then, we study the use of both types of information in a laboratory experiment. Our experimental results show that subjects request both kinds of information evenly, but assign more weight to direct information when deciding whether or not to reciprocate. A decrease in the reliability of indirect information does not affect cooperation levels, but increases the demand for direct information.

This chapter is organized as follows. In section 2 we describe simulation results intended to set a benchmark for observed behavior. Section 3 contains the setup for the experiment, of which the results are presented in section 4. Section 5 provides the conclusions.

2.2 The helping game: simulations

Standard equilibrium theory predicts that in any finitely repeated helping game, backward induction will cause cooperation to unravel. Limited foresight may permit some cooperation in early rounds. The folk theorem on infinitely repeated games applies to both direct and indirect reciprocity and therefore does not give prediction as to which kind may be prevalent.

To get more insight in the basins of attraction of distinct strategies that combine direct and indirect reciprocity, we ran an evolutionary simulation. We investigated the long-term consequences of strategies that employ both IDR and IIR. As a benchmark, we replicated the findings of Nowak and Sigmund (1998), and subsequently extend their model by allowing for strategies that take IDR and combinations of IIR and IDR into account[5]. In the simulations, the history of cooperation by an agent is provided in two formats: the (IIR) image score, quantifying the agent's actions towards others in a range between -5 to 5, and the (IDR) 'memory' of the agent's own interactions with a partner, also ranging from -5 to 5. Each agent carries a heritable strategy *s* denoting the relative weight he assigns to the IDR and IIR scores of a partner. Threshold values *t* that define whom to help ($t \in \{-1, 0, 1\}$) are assigned randomly to agents every new generation. We chose to investigate the evolutionary pressure on *s* separately and reduce the values to arbitrary stable states for *t* to prevent drift.[6] The three threshold values we chose correspond to the long stable phases observed by Nowak and Sigmund in their 1998 paper and cover qualitatively all possible strategies.

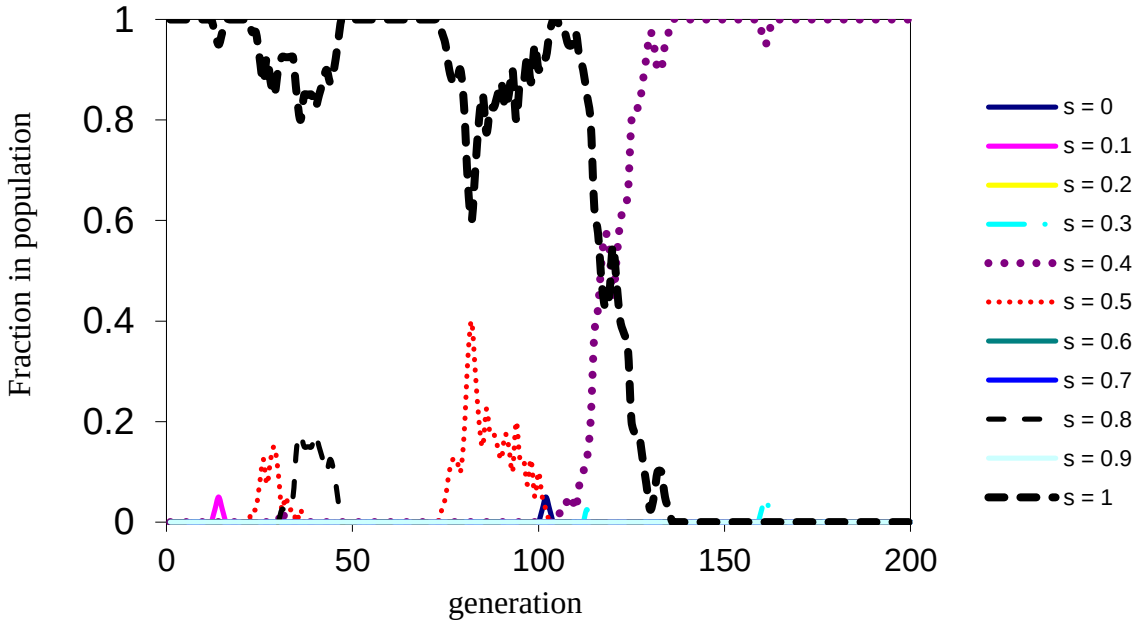
An agent carrying strategy *s* and threshold *t* will offer help if $t \leq s * IDR_{[partner]} + (1 - s) * IIR_{[partner]}$.In words: if an agent's threshold is lower than the weighted sum of his direct and indirect reciprocity information about his partner, he will offer help. Fitness (and subsequently a relatively high share in offspring in the next generation) can be obtained by receiving help from others. For comparability this is set at the same levels as in the Nowak and Sigmund paper: a helping agent incurs a cost of 0.1, the partner receives a benefit of 1. Offspring inherits the strategy of the parent, with a mutation probability equal to 0.025. A decision to

[5] The simulations were run in StarLogo, a Java-based program, version 2.21.
[6] Had we implemented selection pressure on the parameter *t*, too, the model would have converged to arbitrary combinations of *s* and *t*. Alternatively, linking *s* to *t* in any way would impose a constraint.

help leads to an increase in IIR and IDR of 1 point and a decision to pass leads to a decrease of 1. We further assume that information on indirect reciprocity is less perfect than information on direct reciprocity. To implement this, we introduce noise in the IIR measure by introducing a probability of 1/6 that a decision to help decreases the image score, or a decision to pass increases the score. The number of encounters per generation is set at 18000, and the population size at 40, which is the largest number we were able to run on our computers. On average, each individual is paired as donor with any specific other agent 5.77 times. We ran the simulations for 200 generations, which proved to be long enough for convergence.

We first consider a society in which no image scores are formed based on behavior towards third parties but only image scores based on direct information. We did so by initializing all agents such that they only use IDR (s = 1). When doing so, we observed that mutants using a combination of direct and indirect reciprocity information were able to invade the population, although not in every run (see Figure 2.1 for a typical example of an invasion). 7 out of 70 runs converged to a strategy that was a mixture of direct and indirect reciprocity. Although this only makes up ten percent of the total number of runs, the result shows that with our parameters it is in principle a feasible strategy to not only use one's own memory, but also the other agent's image score.

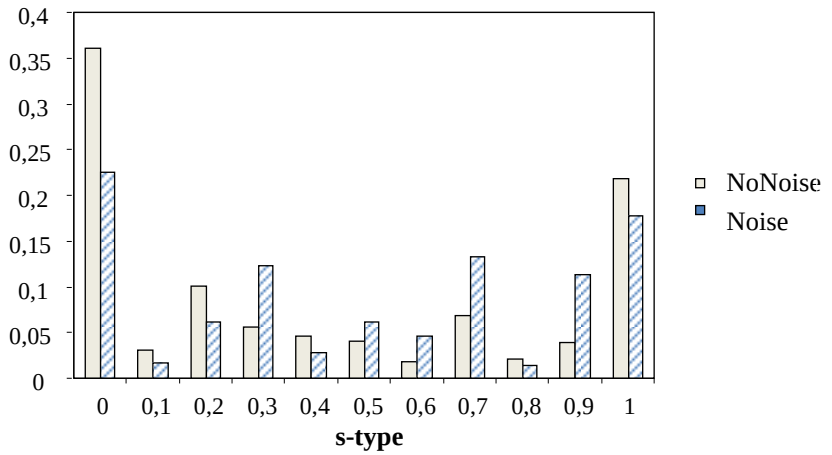
Figure 2.1. Invasion of strategies using reputation



Notes. Lines show the fraction of agents in the population with specific strategies in a typical invasion run over generations. Strategies are defined by the relative weight *(s)* they assign to direct information. The strategy using only IDR (*s*=1, dashed line) loses its dominance and is invaded by a strategy combining IDR and IIR (*s*=0.4, double line). See the main text for a definition of parameters.

As a further test of the stability of strategies that combine information for direct and indirect reciprocity, we initialized the population with randomly picked strategies and ran that simulation 50 times with and 50 times without noise. Cooperation levels (i.e., choices to help) were around 66% (70%) in the sessions with(out) noise. The average *s* (the weight an agent attaches to IDR relative to IIR) of all strategies is 0.40 in the simulations without noise and 0.50 in the noise condition (ANOVA: F=3.854, $d.f.$=1; p=0.05). Hence, noisy IIR information increases the weight attached to DR, but does not eliminate agents using IIR information (see Figure 2.2). Finally, we observed that no single weighted combination of IDR and IIR dominates the runs. Though strategies using both IDR and IIR information survive in this environment, no optimal combination appears to evolve.

Figure 2.2 Influence of noise on strategies



Notes. Fraction of agents with a specific strategies averaged over 50 runs in generation 50-200 with (blue bars) or without noise (red bars). Strategies range from using IDR only (s=1) via a weighted average of IDR and IIR, to strategies using only image scores (s=0).

All in all, our simulations show that the use of image scores can have a selective advantage in a population that can also condition their strategies on personal experience with a partner. Even when indirect information is noisy, strategies that use a combination of both types of information survive evolutionary pressures. Moreover, such strategies are able to invade populations that start off with only strategies solely based on direct experience.

2.3 Experimental design and procedures

A computerized experiment with human subjects was run at the CREED laboratory of the University of Amsterdam, in June and July of 2007. The 120 participants were students from various departments, including economics (43%) and psychology (15%), with an average age of 23. Two treatments were conducted, each with 5 independent groups of 12 subjects, making about 50 decisions each. For every session subjects are randomly assigned to cubicles in the laboratory. No communication among participants is allowed. Written instructions for the experiment (in Dutch; see appendix A for an English translation) are provided. A quiz is used to ensure that the subjects have understood the instructions. When all subjects have finished reading the instructions and have answered the quiz correctly, the experiment starts. Subjects know that after 100 rounds, a next round will start with a probability of 90% (this is done to minimize end game effects). Every session lasts for approximately 90 minutes. Subjects receive, in addition to a show up fee of 7 euros, an initial endowment of 3000 points (300 points = €1). To avoid income effects as much as possible, no information is given about the subjects' current earnings during the experiment, although subjects can calculate these with pen and paper. On average subjects earned €34.50, including the show up fee[7]. At the end of the experiment they are asked to fill out a questionnaire about their personal background and the way they made their decisions.

The setting of the experiment is a helping game. In each round every subject is randomly paired to another; one being assigned the role of donor, the other of recipient. The donor must choose to either pass a specified amount to the recipient, at a cost to himself; or to pass, resulting in no change in payoff for either. Parameters are chosen such that helping costs 150 points to the donor and yields 250 points to the recipient. In the experiment, the choices to help or pass are referred to as yellow and blue choices, respectively. Before the donors decide to help or pass, they are offered the possibility to request information about previous actions by their recipients, when the latter was in the role of donor.

[7] This was the equivalent of $48.

Our experiment closely follows the design in Seinen and Schram (2006), which is a direct translation of the Image Scoring model of Nowak and Sigmund to a laboratory setting. As noted by Bolton et al. (2005), with the group size of fourteen[8] used in the Seinen and Schram experiments, two subjects will 'meet' quite often, though they do so without recognizing each other. Therefore, the Seinen and Schram results on indirect reciprocity may implicitly be blurred by direct reciprocal motivations. In our experiment this re-matching with prior partners was made explicit when presenting information about the current partner's past behavior by distinguishing between his choices when previously facing the donor concerned on the one hand and when facing other parties on the other. Since this experiment was run in the same laboratory as the Seinen and Schram experiments, we will compare our results to their findings instead of running a separate control treatment with only indirect reciprocity[9].

We implement the provision of these two kinds of information as follows. Two boxes are displayed on the donor's computer monitor. Ticking the first gives information about the donor's own experience with this recipient in previous encounters between the two, when the roles were reversed. There is no other way for them to retrieve this information, since subjects cannot be identified. The second box gives the donors insight in the recipient's reputation, by showing choices of this recipient when paired as a donor with others. If requested, the direct and indirect information summarizes the previous six decisions of the recipient in the role of donor[10]. The donor's own public score may be computed, but is not presented. When requested, only the *number* of the current recipient's blue and yellow choices in the last six decisions is given, not the order. This limited information reflects limited memory and gives subjects an opportunity to clean their record. Note that it takes longer to gather direct than indirect information about the helping behavior of another subject. Therefore, for a long part of the experiment (on average 66 rounds), the indirect information is based on more observations than the direct information. Avoiding this problem, which reflects a trade-off that we also face in real life, would have required either a very small group size or an even larger number of rounds.

[8] We used a group size of twelve.
[9] But see Molleman et al. (2013) for the comparison with another control treatment.
[10] In early rounds, the total number of decisions may be smaller than six, or even zero.

We impose a cost of five points for every information request, to ensure that subjects deliberately click on the information they are interested in[11]. Note that these costs are very low compared to the costs of helping (150) or the benefits of receiving help (250). We did not impose an order in the information requests, since we are interested in which type of information people prefer when both types are available. After the decision has been made, the donors and recipients are informed about their earnings in the round.

To study the extent to which subjects are sensitive to the reliability of images we implemented two treatments similar to the simulations in section 2. In the *baseline* treatment the information on both direct and indirect information is completely accurate. The alternative, *noise*, treatment deviates only in the reliability of the indirect information; one out of every six pieces of information is switched from positive to negative or vice versa. This intends to reflect distortions of information that may occur, for example, as a consequence of gossip. This implementation is equivalent to a noise level of 17 % and does not affect the standard equilibrium predictions as noted in section 2. Note that our implementation is slightly biased against extreme scores, since an observed 5:1 score is more likely to stem from a true 6:0 than from a true 4:2, but since these extreme values (0 times uninterrupted helping or passing) hardly occurred, this effect is unlikely to have affected our results, should any of the participants have noticed.

[11] An alternative would be to let the cost depend on the amount of information they receive, but that would make it more difficult to infer a person's preference for each type of information.

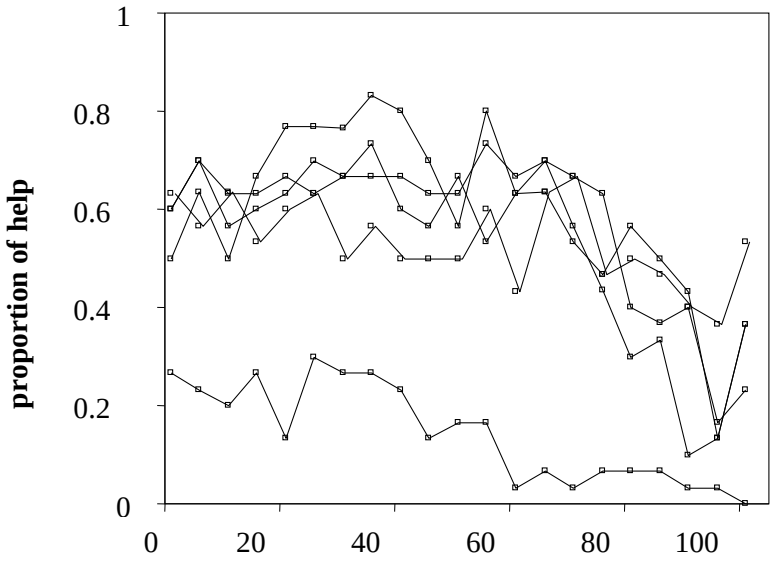
2.4 Experimental results

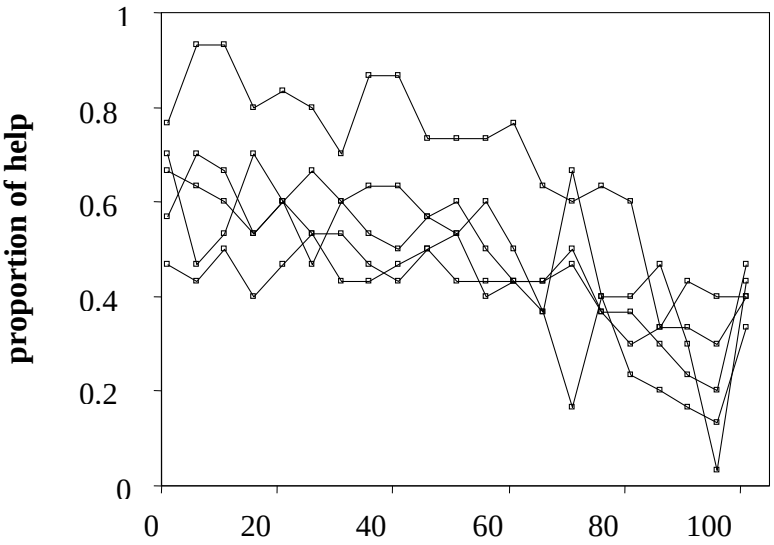
This section is divided into four subsections. After presenting overall helping levels in 4.1, the next subsection gives an overview of the information requests (4.2). The aggregate influence of information on helping behavior is discussed in 4.3 and followed by a detailed overview of individual types (4.4).

2.4.1 Helping levels

Help was given in 49 % of all interactions. Reciprocity is a strong phenomenon in our setting: the correlation between the donors' cooperative choices and the number of times they received help previously by others is very high (Spearman $r = 0.88$, $p<0.001$). Note that this correlation does not express the direction of help: it may be due to direct (A helps B, then B helps A), indirect (A helps B, then C helps A), or generalized (A helps B, B helps C) reciprocity. The average helping level in the first 75 rounds of our baseline treatment (57%) is comparable to the level found in a similar treatment by Seinen and Schram (70%), where information about the six last choices was provided (costlessly) in every round. Furthermore, we observe that the average cooperation level declines, notably so after round 75 (see Figure 2.3).

Figure 2.3. Helping levels





Notes. Average proportion of helping per group per 5 rounds in the baseline (upper panel) and the noise (lower panel) treatment.

In the first 75 rounds the fraction of helpful choices in groups was on average 0.56 (*s.e.* 0.05); we do not find a significant difference between the treatments without (0.57) and with (0.53) noise (MW (10 groups) z=0.629, p=0.53). Within treatments, however, large differences are observed across groups. In particular, one outlier group in the baseline treatment shows a distinct pattern of only 19 % cooperative choices (see the upper panel in figure 2.3). Closer inspection of the individual data reveals five of 12 subjects in that specific group who never helped.

These results on helping choices are summarized by:

Result 1: *Helping levels are high and decline over rounds. In aggregate, there is no difference in helping between the baseline treatment and the noise treatment.*

2.4.2 Information requests

In on average 49% of all rounds subjects used the possibility to request costly information about others' previous choices. Table 2.1 provides an overview of the information requests. Donors requested (direct) information on decisions that concerned behavior of the recipient towards themselves (32% of the interactions) and indirect information on decisions that concerned others (26%). Without noise the frequency of requests for these two kinds of information was almost equal (29%, 28%). In the noise treatment, the direct information was requested more often (36%: 24%), but not significantly so (Wilcoxon $T = 2$, $p = 0.14$, r=0.44, N = 10). The difference in IDR requests between treatments is not significant either (MW (5 groups) z=0.73, p=0.46), nor is the difference in IIR requests (MW (5 groups) z=0.73, p=0.46).

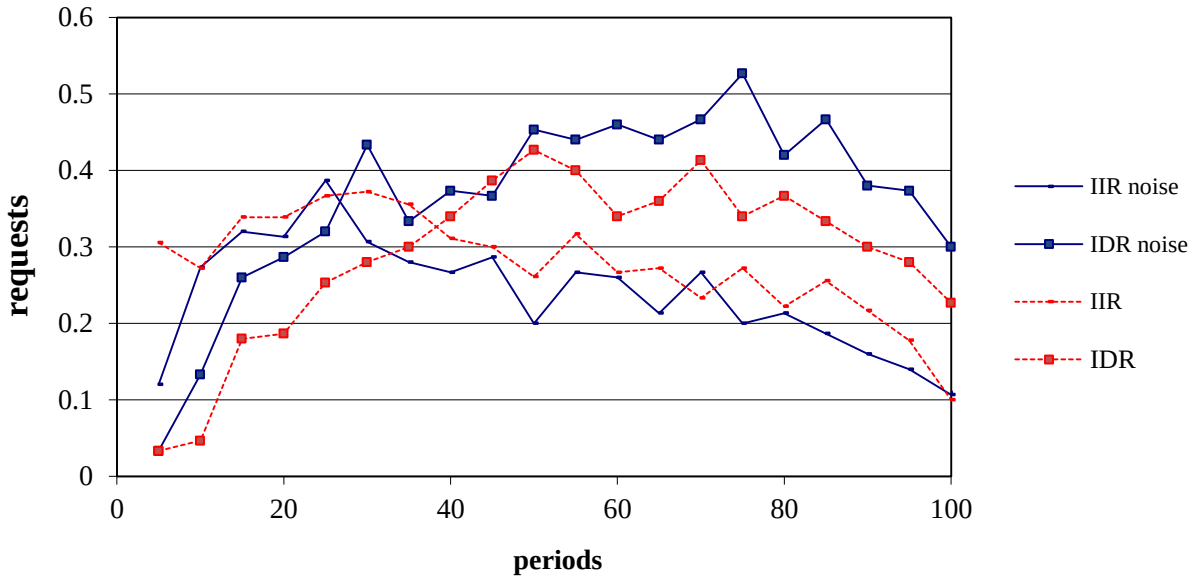
Table 2.1. Information requests

	IIR and/or IDR	IIR	IDR	IIR and IDR
Baseline	46%	28%	29%	11%
Noise	52%	24%	36%	8%

Notes. Frequency of information requests (any, information for indirect (IIR) or direct (IDR) reciprocity or both) for the baseline and the noise treatments.

The development of information requests across rounds is shown in figure 2.4. It appears that the total number of IDR requests increased slightly across rounds until round 75 (Spearman r = 0.04, p<0.05). IDR becomes more popular at the expense of IIR (r = 0.14, p <0.001); from round 40 onwards, IDR was requested more often than IIR in both treatments (Figure 2.4). This may indicate that the subjects understood the incremental nature of this information, and specifically the fact that gathering direct information takes more rounds than indirect information. The difference between IDR and IIR requests is significant from round 45 on in the noise treatment and marginally so in the baseline treatment (noise: MW (5 groups) z=1.98; p=0.04; baseline: MW (5 groups) z=1.77, p=0.07).

Figure 2.4. Information requests over rounds



Notes. Fraction of subjects per round that requested information on direct (IDR) and indirect reciprocity (IIR) for the baseline treatment (in red) and the noise treatment (in blue).

We again summarize these findings.

Result 2: *The number of requests for direct reciprocity information increases over the first 75 rounds and is higher in the noise treatment. The number of requests for indirect reciprocity information decreases after the first 25 rounds and is lower in the noise treatment.*

2.4.3 Use of information

The information provided is a score either summarizing the recipient's choices in previous encounters with the donor (information for direct reciprocity) or others (image score). We classify a score as positive if the observed helping choices outnumber or are equal to the observed choices to pass[12]. Subjects reacted strongly to positive and negative information; see Table2.2 for the percentage of decisions to help subjects with a positive or negative score.

[12] We classify 'no information available' as positive.

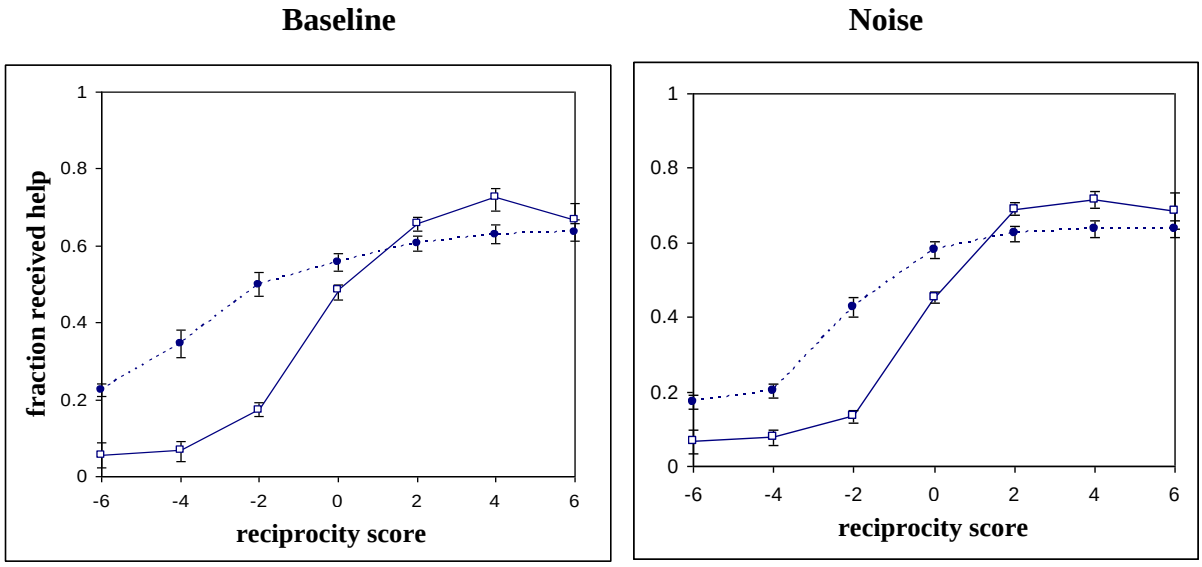
Table 2.2. Helping behavior following information

	Baseline		Noise	
	Positive score	**Negative score**	**Positive score**	**Negative score**
IDR exclusively	85 %	10 %	81 %	4 %
IIR exclusively	85 %	18 %	70 %	32 %
IDR and IIR	86%	10 %	92 %	0 %
No info requested	50 %		35 %	

Notes. Percentage of decisions to help after requesting direct (IDR) and/or indirect information (IIR) and learning the outcome in the baseline and the noise treatment. See Table 3.5 for the cases where both IIR and IDR had been requested, but yielded conflicting information (not included here).

In the treatment without noise, subjects were much less inclined to help after receiving negative information, both in the case of indirect (18% cooperation) and direct information (10% cooperation). Positive information led to help in 85% of the encounters, both for direct and indirect information. When no information had been requested help was given in 50% of the cases, perhaps indicating subjects' strategic concern for their own score. This is corroborated by the finding that subjects sometimes offered help even after receiving negative information about the recipient. Uninformed decisions were on average less cooperative in the noise treatment (35%), which could be explained by subjects' hoping to 'hide behind the noise'. As a general tendency noise dampens the effect of information by adjusting helping rates after positive or negative IIR to the average (positive: from 85 to 70%; negative: from 18 to 32%).

Figure 2.5. Average receiving rates per score



Notes. Lines show average receiving rate for direct (full line) and indirect reciprocity scores (dotted line) in the baseline and the noise treatment.

The expected help as a function of a subject's score is shown in Figure 2.5 [13]. The graphs show a sharp increase between a direct score of -2 and 0, both in the baseline and the noise treatment[14]. In comparison, the graphs for indirect information are flatter. The marginal effect of a higher indirect score is more constant. Having a negative direct score thus reduces the receiving rate more than a negative indirect score does. A possible explanation is that subjects prefer to use direct information, if available, and use indirect information in case of doubt.

To correct for other factors that may contribute to the decision to help, such as the received help so far, we present a probit regression with random effects at the group level (to correct for interdependencies within groups). We estimated two models; both including observed scores as independent variables. One also includes the events of the previous round and the other includes average variables up to round *t*. Table 2.3 shows the results.

[13] In 10 % of the cases subjects requested both types of information. Since we do not know on which information they conditioned their decision, we include all observed and unobserved scores in the graphs instead of counting some decisions twice. Removing the decisions that were based on unobserved scores or on both types of information does not change the graph qualitatively.

[14] To obtain more reliable data (i.e., sufficient observations with certain scores) we aggregate by rounding up uneven scores to the next even integer.

Table 2.3. The decision to help

Dep. var: Help of *j* by *i* in *t*	**I: lagged help**		**II: average help**	
Treatment	**Baseline**	**Noise**	**Baseline**	**Noise**
Observed IDR score in *t-1*	0.45***	0.57***	0.42***	0.53***
Observed IIR score in *t-1*	0.21***	0.24***	0.21***	0.24***
Round	0.41	1.15***	0.64	1.31**
Round^2	-1.31***	-2.18	-1.90	-2.74
Received help *t-1*	0.34***	0.40***		
Gave help *t-1*	1.04***	0.75***		
Average help *i* received			-0.03	0.42**
Average help *i* gave			2.85***	2.31***
Constant	-0.56***	-0.61***	-1.49***	-1.44***
# observations (groups)	2974 (5)	2963 (5)	2974 (5)	2963 (5)
log likelihood	-1343.56	-1256.75	-1152.18	-1098.82

Notes. The table presents the results of a random effects probit regression used to explain help of j by i in t. Formally, it gives the estimated coefficient vector β in $Pr(Help)_{jit} = \Phi(\sum i\ X'_{ijt}\beta + \mu_m)$ where $Pr(Help)_{jit}$ is the probability that i helps j in round t; Φ denotes the cumulative normal distribution and X_{ijt} is a vector of variables relating to i and j in t as described in the first column of the table. μ_m is a (white noise) matching-group-specific error that corrects for the dependencies within matching groups. Variables included are the observed IDR score, the observed IIR score, round, round squared, help received in the previous round as a receiver, help given in the previous round as a donor, IDR IIR cumulative average help received up to t, cumulative average help given up to t, and a constant. The first line per variable denotes the coefficient (*, **, *** indicating significance at the 0.10, 0.05, and 0.01 level, respectively).

As Table 2.3 shows, both the observed direct and indirect score are highly significantly positive in both models. Other things equal, an increase in the observed IDR score increases the propensity to give help more than an increase in the observed IIR does in both treatments (z=7.42, p<0.001; z=9.81, p<0.001; marginal effects: observed IDR = 0.17, observed IIR = 0.09 (baseline): observed IDR = 0.21, observed IIR= 0.09 (noise)), corroborating Result 2

about IDR being preferred; The noise treatment further increases the difference between the impact of IDR and IIR.

The distinction between models I and II lies in how the donor's history is added to the model. These models allow us to test for the effect of generalized reciprocity (if A helped B, B helps C). In model I, history is restricted to one round. We include dummies for having received or given help at the previous respective opportunity[15]. The coefficients for both are significantly positive. The first suggests that in both treatments subjects are motivated by some form of non-strategic helping (i.e., generalized reciprocity). The coefficient for the dummy for having helped in the previous round picks up individual differences in tendencies to help.

The second model confirms that the effects of the observed scores are robust against a control for histories, now represented by the average across all previous rounds. The most remarkable finding is the large and significantly positive influence of average helping rates in both treatments. This strongly confirms the individual heterogeneity in the propensity to help, even after controlling for what donors observe and how much help they have previously received on average. A second noteworthy difference between the two models is that the average help now received has no significant influence on helping in the baseline treatment. This is an indication that generalized reciprocity has a limited memory in the sense that only recent good experiences are positively responded to. Although being helped in the previous round significantly increases the propensity to help, on average generalized reciprocity cannot explain the help given in the baseline treatment.

This brings us to the third main result.

Result 3: *The direct reciprocity score has more impact on the helping decision than the indirect reciprocity score. Noise on indirect reciprocity information diminishes the relative impact of this information on the helping rate.*

[15] We tested for the effects of imbalances between the number of times an individual has been in the role of donor compared to receiver, but found only insignificant or negligibly small (0.001 smaller than other) effects.

2.4.4 Individual strategies

We can define a donor's strategy as a mapping from the information seen (and possible the own reputation) to a choice whether or not to help. We therefore classify individuals based on their information requests. This yields four types: those who did not request any type of information more than twice; those who requested both types of information three times or more; and subjects who asked for either direct or for indirect reciprocity information more than twice, but not for the other type[16]. Below, we provide an overview of the behavior per type (see Table 2.4). We will discuss our observations for each type in turn.

Table 2.4. Behavior per type

Type	**No Info**		**Both Info**		**DR**		**IR**	
	Base	**Noise**	**Base**	**Noise**	**Base**	**Noise**	**Base**	**Noise**
% subjects	27	23	43	42	12	20	18	15
Helping rate (%)	26	19	62	65	70	47	46	48
No. requests IDR	1	0	52	59	51	57	1	1
No. requests IR	2	1	41	36	1	1	50	57
Av. earnings	6319	5480	5517	6041	6174	5818	6160	5540

Notes. Rows give, respectively, the player type, the percentage of subjects classified as a certain type, their average helping rates, average number of requests for information for direct and indirect reciprocity and earnings for the two treatments.

No Info. Subjects who never requested any information make up for a quarter of the participants (baseline: 27%; noise: 23%). Since their decisions are not influenced by information (except for their personal record of receiving help), we can only analyze the general 'giving' pattern in this group. On average they help in 26 (noise: 19) % of the interactions; 15 of the 30 subjects of this type never helped, 5 subjects behave as unconditional

[16] Since subjects requested both types of information in 10% out of 50 rounds on average, we chose to set a cut-off between types at two requests or more.

cooperators and helped in (almost) every interaction, and the others in this category alternated between 'giving' and 'passing' in some orderly manner[17].

Both Info. The second class of subjects consists of those who regularly (more than twice) requested information about both types of information. This is the most common type (baseline: 43%; noise: 42%). Subjects in this category helped in 63% of all encounters. They requested IDR in 56 %, IIR in 39% of the rounds, and both in 9%. In 20.1% of these cases (165 observations) the two sources of information conflicted, i.e., the sign of the image score in the indirect information did not correspond to the sign of the score in the direct information. Table 2.5 shows choices for this group of donors. Although this is a relatively small sample of all observations, it reveals a similar preference for direct reciprocity as the information requests pattern: in 63 % of these conflicting cases people make a decision in line with the direct information they received. Noise changes this pattern slightly (and counter-intuitively) towards more indirect information, but not significantly so.

Table 2.5. Conflicting information

	Baseline	Noise	Overall
Subjects behaved in line with IDR	69 %	60 %	63 %
Subjects behaved in line with IIR	31 %	40 %	37 %

Notes. Percentage of decisions that were in line with either IDR or IIR when the two scores were conflicting.

IDR. The third category consists of subjects who mostly requested direct reciprocity information (on average in 56 % of the rounds), but hardly ever requested indirect information. There is no significant treatment effect on the proportion of these types (MW, N=20, z=0.86, p=0.39); these IDR types offered help more often in the baseline treatment than in noise (70% vs. 47%) but not significantly so at the group level (MW, N=8, z=1.64, p=0.10).

[17] One of them exercised a distinct repeated pattern consisting of one time 'passing', two times 'helping'- an example of the 'score optimisers' who cares only about their own score (Engelmann and Fischbacher, 2009). These subjects earned 1.67 times the average, confirming that such strategic types can exploit image scorers (Leimar and Hammerstein 2001).

IIR. The subjects in the final category mainly asked for indirect reciprocity information. They helped in 47 % of the cases. This is less than the IDR types. Since the average of the two scores follow the same pattern, this suggests that IIR subjects require a higher score before they give help. We do not observe a treatment effect for either the proportion of subjects that fall into this category or the helping rates.

Finally, earnings do not differ significantly across types. This can be compared to the results from the evolutionary simulation, where we do not observe any single strategy dominating others.

Our results are summarized as follows.

Result 4: *There is heterogeneity in behavior with respect to preferences for and reactions to reputation and direct experience. The types do not differ with respect to earnings.*

2.5 Discussion

This is to our knowledge the first experimental study in which people are given the choice between direct and indirect information about their partner's history of cooperative behavior. In this chapter we have shown that humans use both direct and indirect information when deciding about a reciprocal gift, even when the indirect information is less reliable. Summarizing, we find that 1. helping levels do not decrease in a noisy environment; 2. people substitute more reliable direct information for noisy indirect informatione; 3. direct information has more impact on the decision to help on average, but 4. people consistently differ in their use of the two types of information. Each of these results is put into perspective below, followed by a reflection on the link between the results from simulations and experiments.

Previous studies provide all participants with information about aggregated previous behavior (Seinen and Schram, 2006; Engelmann and Fischbacher, 2009). In comparison with the findings of Seinen and Schram (2006), information in our setup came at a cost, yet helping rates appeared to be only slightly lower. This shows that people are willing to incur a cost for either type of information. It has been argued that indirect reciprocity is likely to dilute in large groups because of the noisy nature of information (Engelmann and Fischbacher, 2009). We find, in contrast, that people show sensitivity to noise not by lowering their propensity to help but by switching to their own experience. This demonstrates that people can cope with relatively unreliable gossip and respond by switching to more reliable sources of information.

As discussed in our overview of the literature, some studies have shown that behavior in an indirect setting affects choices in a subsequent direct setting and vice versa, suggesting that the two forms of reciprocity are exchangeable (Wedekind and Braithwaite 2002; Milinski et al., 2002). These findings reverberate with our observation that people substitute direct information for noisy indirect information. Our results provide direct evidence that people integrate indirect information with their own experience.

To date only a few studies have investigated whether people react differently to direct and indirect information about others' previous actions[18]. Studies that attempt to tease apart the effects of direct and indirect information on helping report inconclusive results (Dufwenberg et al. 2001; Bolton et al., 2005). We observe that direct reciprocity is more often decisive for the decision to help, but we find no difference in average donation rates between users of direct and indirect reciprocity information. Neither do we find that helping rates immediately after receiving positive direct information differ significantly from helping rates after receiving positive indirect information; nor do helping rates differ after negative direct versus negative indirect information. Only when we compare direct information to noisy indirect information do we find a difference in helping: after positive (negative) noisy information we observe less (more) helping than after direct information. This dampening effect of noisy indirect information on helping rates can be compared to the observation in Sommerfeld et al. (2007), who find that third party information (which was not necessarily accurate) leads to a less pronounced reaction to positive or negative behavior than the direct observation of that behavior.

Like in real life situations, most participants in our experiment use indirect information alongside direct information; those who do not combine them show a clear preference for either of the two. The heterogeneity in strategies we observe does not lead to large differences in earnings. The finding that people differ consistently in how they use and weigh the two kinds of reciprocal information suggests that studies that do not distinguish between the two risk overlooking structural behavioral patterns (cf. the literature on personalities in behavioral ecology, Wolf et al. (2008)).

The use of direct reciprocity has been established in various species (Dugatkin, 2002) and some species appear to use mental bookkeeping in order to direct help towards individuals that have been helpful to them (Krams et al. 2008, Wilkinson 1984). Indirect reciprocity, on the other hand, has only occasionally been documented in other species (Jansen and Van Baalen, 2006), but is widespread among humans (Nowak and Sigmund 2005). Models taking direct

[18] Van den Broek and Hopfensitz (2010, working paper) disentangle the effects of emotions evoked by direct experience versus third party reputation information in a repeated trust game.

and indirect information into account have thus far been restricted to strategies using either of the two (Roberts, 2008). Our simulations suggest that strategies incorporating both types of information are key to the invasion of a strategy using indirect reciprocity information. Although simulations are sensitive to parameter settings and the specific properties of the implementation, these predictions were echoed by the experimental findings. Broadly speaking, they suggest that the use of indirect information in addition to direct information can evolve and be maintained, even when indirect information is less reliable than direct information and all other agents use only direct information. Indeed, we observe distinct self-consistent strategies in the human population that do not yield differences in earnings. These behavioral results suggest that there is not much evolutionary pressure on individual variation in preferences for direct and indirect reciprocity information.[19] Together, these results may contribute to the long-standing issue of how the use of reputations has gained a foothold in social interactions. The combination of simulation and experimental results indicate that people use costly information on noisy indirect interactions in addition to information on direct interactions, and that it is adaptive to do so.

[19] An alternative explanation would be that disruptive selection causes different personalities to arise that coexist in a polymorphism (Wolf et al., 2011).