Trust and trustworthiness reputations in an investment game

Gary Charness [a,*], Ninghua Du [b,1], Chun-Lei Yang [c,2]

[a] *Department of Economics, University of California Santa Barbara, Santa Barbara, CA 93106-9210, United States*
[b] *School of Economics, Shanghai University of Finance and Economics, Shanghai 200433, PR China*
[c] *Research Center for Humanity and Social Sciences, Academia Sinica, Taipei 115, Taiwan, Republic of China*

A R T I C L E   I N F O

A B S T R A C T

Trust is an essential component of good social outcomes and effective economic performance. Reputation on the trustee's past behavior in the same role has proven to be greatly effective at raising the level of trust. In this study, we show that providing information on the trustee's past behavior as the trustor is equally effective as a reputation system. In fact, people still find it worthwhile to invest in a reputation as a trusting person, even though the immediate payoff for trusting is poor. This confirms the role of *indirect reciprocity* as a strategic notion based on reputation, whereby pro-social actions by one person towards a second person are sanctioned by a third party.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

A major concern in social-dilemma games is how to exploit gains from trade that are potentially achievable within some form of social interaction, but which are hard to realize due to individual incentives to maximize own profits. Pro-social behavior in social dilemmas is crucial for economic performance, whether as social capital or as a lubricant. In abstract terms, the social optimal outcome may not be part of any equilibrium or be only one of many, most of which are inefficient. Typical environments include the prisoner's dilemma, public-goods games, one-side giving, and the investment ('trust') game.

The concept of trust has taken a central position in economics and social sciences alike. Fukuyama (1995) presents a theory of trust in which a country's economic prosperity is correlated with the amount of social capital existing within that country. The concept of social capital, described by Coleman (1990), refers to the features of social organization (including trust and trustworthiness) that improve efficiency by facilitating cooperative actions. Ostrom et al. (1992) find that people tend to trust one another to adhere to nonbinding agreements when given an opportunity to communicate and to choose

* Corresponding author. Fax: +1 805 893 8830.
  E-mail addresses: charness@econ.ucsb.edu (G. Charness), ninghua.du@mail.shufe.edu.cn (N. Du), cly@gate.sinica.edu.tw (C.-L. Yang).
  URLs: http://www.econ.ucsb.edu/~charness (G. Charness), http://iclass.shufe.edu.cn/teacherweb/users/ninghua.du/ (N. Du), http://idv.sinica.edu.tw/cly/
(C.-L. Yang).
  [1] Fax: +86 21 65904198.
  [2] Fax: +886 2 27854160.

their own property rights, while Ostrom and Walker (2002) examine the importance of reciprocal relationships in explaining the origins of trust and trustworthy behavior.

In this regard, a positive note is that the familiar folk theorem tells us that cooperative behavior can be sustained with infinite repetition and sufficiently patient players. Indeed, this may be a fairly close match for small societies where repeated and non-anonymous interaction is quite frequent, so that socially-beneficial behavior can develop; in a sense, one's reputation may augment any pure altruistic tendencies. However, in a larger society this system is likely to break down without further intervention. Since contemporary interactions (such as web-based auctions and e-commerce) are frequently not face-to-face, but are instead anonymous and one-shot affairs, it is critical to devise mechanisms to prevent such breakdowns in cooperative behavior.

One distinction that has been made in the literature is concerned with whether a choice is motivated by something ingrained in an individual's social preferences or by strategic reputation-building. The presence of social preferences makes it easier to jump-start the reputation system, as well as to help in stabilizing it. However, the strategic element is important for making the system self-sustaining.[3] While reciprocity is one form of relevant social preference, it is not self-evident that a reputation system for indirect reciprocity may function at all well.

Evolutionary biologists such as Nowak and Sigmund (2005) distinguish between direct and indirect reciprocity.[4] The latter is considered to be a strategic notion based on reputation and to be sustainable in an evolutionary sense. Panchanathan and Boyd (2004) directly link indirect reciprocity to a reputation for cooperative behavior (with indirect reciprocity breaking down without such a link), and Fehr (2004) suggests that individual reputation is key to providing a suitable evolutionary explanation for the survival of collective action in large groups. For reputation to have an effect, it seems clear that people must believe that they will be rewarded for having a good reputation, so that it becomes worthwhile in the long run to invest in a good reputation by making an immediate sacrifice in payoffs.[5,6]

Given the difficulties involved in making accurate assessments about the degree to which indirect reciprocity and reputation can succeed in promoting pro-social behavior in the field environment, recent experimental work has begun to investigate this question with controlled laboratory studies. While we review this work in more detail in the next section of the paper, we mention here that these previous studies have provided evidence that (different) mechanisms that provide information either about whether an individual has 'helped' (in an allocation task) or has been 'trustworthy' (in a response in a game) is successful in inducing better social outcomes.[7]

The mechanism of a reputation for trustworthiness seems *ex ante* rather likely to sustain cooperative behavior, since even a selfish participant C who observes that person A has been helpful in the past to person B is likely to help or trust person A. In comparison, an indirect-reciprocity mechanism is based on A rewarding B in the hope of being rewarded by C (who has no option of being rewarded by A in turn), so A simply builds a reputation for being 'nice' or trusting. While both of these mechanisms require an information-sharing system to work, one question is whether a trait of being trusting is sufficient to sustain a successful and socially-efficient indirect-reciprocity reputation system in a strategic environment similar to the well-known 'trust' (investment) game (Berg et al., 1995).

We investigate the relative effectiveness of providing information on an individual's history of being willing to 'trust' another party, forgoing an immediate gain to yield a much higher social payoff with a risk of receiving nothing. While it is relatively easy for a participant to see the long-run value in having a good reputation for being 'trustworthy,' it takes more of a leap of faith to believe that others will trust you because you have a history of trusting others, particularly when there are strategic considerations about whether to trust. In this case, the reputation-building sacrifice of the trusting act can be severely dampened by a lack of direct reciprocal response. The instinctive reaction to this negative experience is to withhold trust, which works against the dynamics of building a good reputation. Only when a substantial proportion of people believe in and work towards reputation building can we expect that it is worthwhile for them to continue to do so, perhaps attracting others to do the same.[8]

---

[3] Note that Engelmann and Fischbacher (2009) find while there is evidence of "pure" indirect reciprocity, strategic reputation-building options play a role. Also, Yang et al. (2006) show that people's inclination for pro-social actions is closely correlated with the monetary bargaining strength associated with their respective roles in a large range of games.

[4] The term indirect reciprocity was introduced in Alexander (1987). Nowak and Sigmund (2005) remark on p. 1291 that: "The evolution of cooperation by indirect reciprocity leads to reputation building."

[5] The theoretical work on the evolution of indirect reciprocity typically requires higher-order information. For example, in the context of the prisoner's dilemma, an individual may have chosen to defect as a punishment for the other party's having defected in the past. This pro-social action is indistinguishable from selfish defection unless the history of the other parties who have been matched with the individual (and so on, recursively) is provided. For simplicity, we elect not to deal with this in the current study and refer to Gong and Yang (2010) for an explicit design with second-order information.

[6] In fact, the need for information about the past behavior of strangers has spawned a variety of reputation systems. One such system is that of the credit bureaus, who provide information about an entity's payment history. More recently, online e-commerce sites such as Amazon and eBay allow feedback from parties to a transaction, with this feedback aggregated for each online entity and available for inspection by all parties interested in doing business with this entity. The extent to which such reputation systems promote cooperative behavior in the field remains an open question, as this is currently evolving. Nevertheless, a study by GartnerG2 (2002) illustrates the potential severity of the problem, as "Internet transaction fraud is 12 times higher than in-store fraud."

[7] Although we recognize that other motivations (such as social preferences, discussed later in the paper) are likely to influence behavior, for the sake of simplicity we nevertheless use the terms trust and trustworthy to reflect, respectively, whether the first mover has foregone her outside option and whether the responder has selected the option that benefits the first mover.

[8] In a sense, it is like a network effect where the marginal benefit of partaking in the network increases with its size.

We have three treatments in our experiment. In our baseline treatment, no information is provided about an individual's history. In the "History of Return" treatment, the first mover has the opportunity to learn the potential responder's history of making favorable responses. In the main treatment, "History of Trust," a first mover has the opportunity to learn the potential responder's history of trusting responders.

We confirm the results of previous studies in that providing information about a potential responder's history as a responder leads to far more trust than when no such history is available. However, it is something of a surprise that providing information about previous trusting behavior leads to the *identical* overall increase in trust as does providing information about trustworthy behavior.[9] This suggests that the scope for reputation mechanisms may be broader than has been previously thought. We also investigate other determinants of trusting and trustworthy behavior.

On a deeper level, there is a question of precisely why this reputation system is effective. One possibility is that the history of trust could serve as a signal of trustworthiness; indeed, these traits may be correlated well. Another possibility is that trust-as-a-proxy-for-trustworthiness in the early periods is supplemented by a gradual transition to behavior driven by indirect reciprocity in the later stages of the session. We discuss this in some detail later in the paper, concluding that the latter explanation seems to fit better with our data.

The remainder of this paper is structured as follows. Section 2 offers a review of the previous experimental literature on indirect reciprocity, while Section 3 presents the detail of our experimental design. We present our results in Section 4, and provide some discussion in Section 5; we conclude in Section 6.

2. Indirect reciprocity and related experimental literature

2.1. Indirect reciprocity

Ohtsuke et al. (2009, p. 79) state that with indirect reciprocity: "I help you, and someone helps me. Indirect reciprocity is based on reputation. Helping others establishes the reputation of being a helpful individual." In view of this concept, a standard game used in this literature is "The Helping Game," in which only the would-be helper makes a decision in any given period. He or she chooses whether or not to help another party at a cost that is smaller than the benefit for the other party. Information concerning the historical helping behavior of the other party is provided.

With different forms of information provision, indirect reciprocity may work via different modes of interaction between information processing and individual strategies. In the evolutionary literature, the primary concern is which reputation 'score' might work best or has the chance to be part of a stable system. The *image score* is the number of times an agent helped. Nowak and Sigmund (1998) showed through simulation that the whole population evolves to the most discriminating threshold strategy, implying a norm where someone is helped, if she helped others in at least half of the past opportunities. However, Leimar and Hammerstein (2001) show that strategic players, concerned only about keeping an own score that gives a high probability of being helped, can successfully invade a population of threshold strategies.

The so-called 'standing strategy' (Sugden, 1986), where the image score is modified to award someone who punishes a miscreant a positive score instead of a negative one, proved to be much less prone to such an invasion. Another reputation score of importance is "judging," which further modifies standing by giving a negative score to someone who didn't punish the bad guy in the eyes of the system.[10] Nowak and Sigmund (2005) summarized further studies that favor different score-notions as the underpinning of an evolutionarily-stable reputation system that facilitates the proliferation of pro-social behavior.[11]

Note that the helping game differs from the 'trust' (investment) game in one critical aspect, namely that the recipient in the latter has the additional option to make a transfer back to the donor/investor. This generates interactions with direct reciprocity, and a complex theory of co-evolution of both direct and indirect reciprocity would be extremely difficult to develop. However, if we assume a norm among the recipients to never return anything reciprocally, then the system is reduced to an equivalent of the helping game. Thus, under this assumption, all the theoretical predictions developed from the indirect reciprocity literature for the helping game apply in a straightforward manner to the trust game. In fact, to the extent that expectation of direct reciprocity may hinder the proliferation of an indirect-reciprocity norm, the trust game provides a kind of robustness test to see whether indirect reciprocity may prevail against the noise introduced in this return option for the recipient.

---

[9] However, since the trust history has a direct effect on trust (though based on indirect reciprocity), whereas the return history can only work indirectly, it might not be so surprising that the history of trust system works equally well on trust. Nevertheless, the expectation of indirect reciprocity is weaker than the expectation of trustworthiness based on the history of trustworthiness, so the overall effect is not obvious.

[10] As both standing and judging involve higher-order information about the opponent's earlier partners scores, and thus actions, they are difficult to implement in experimental investigations such as ours. Bolton et al. (2005) and Milinski et al. (2001) are exceptions that provide second-order information. The former provide only one-period data and find that this information does play a role. The latter provide all periods of second-order information, but could only observe that it took subjects a lot longer to make their decisions without any evidence of how it enters the decisions.

[11] Note that indirect punishment can serve as an alternative to indirect rewarding, beyond the implicit punishing effect of withholding help. Ohtsuki et al. (2009) show, however, that due to its additional cost it has only marginal chances to prevail evolutionarily. Experimental evidence supports this view. For example, Rockenbach and Milinski (2006) only find a reduced role for third-party punishment in public-good experiments.

2.2. Related experimental literature

Perhaps the earliest experimental study to address indirect reciprocity is that of Kahneman et al. (1986). People choose whether to split $20 evenly with an anonymous second party or to take $18. After a fraction of these choices were randomly implemented, people whose choices were not implemented were arranged in three-person groups, and one person in each group was informed about the earlier choices of the other two people in the group. If these other people had made different choices, the decider was then asked to choose between (Self, Equal Chooser, Self-favoring Chooser) payoffs of (6, 0, 6) or (5, 5, 0). Many participants made the latter choice, sacrificing to either punish an unfair allocator or to reward a fair one (or both); there was also substantial correlation between the choices made in the two stages. While this study employs a rather limited reputation system, it demonstrates a form of indirect reciprocity.

Seinen and Schram (2006) use the Nowak and Sigmund's (1998) image-scoring game, in which the first mover chooses whether to give or not; roles are re-drawn every period. If the first mover gives, this benefits the paired recipient (with the benefit larger than the cost of giving). The cost of giving was varied, but was always less than the benefit to the other party. In one of the high-cost treatments, no history was provided, while in the other two treatments, the donor learns about the previous six decisions made by the recipient when he was a donor. The giving rate was much higher in both the high-cost and low-cost treatments than in the treatment with no information. The giving rate was increasing in the number of helpful choices made by the recipient (when this was shown), in support of indirect reciprocity.

Bolton et al. (2005) also use the Nowak and Sigmund's (1998) image-scoring game. They vary the cost of giving (with the benefit always larger than the cost of giving), as well as the information provided to the prospective donor about the recipient's history of giving. There is considerably more giving when the cost is low; the giving rate increases in the amount of information provided when the cost is high, but does not vary greatly when the cost of giving is low. Crucially, the use of the information there is in line with standing.[12]

Greiner and Levati (2005) study indirect reciprocity in a closed ring of either three or six people; each amount sent to the next person in the ring is tripled. They also vary the matching condition and whether players make simultaneous or sequential choices.[13] The amount a player sends is sensitive to the amount she receives, and there is more sent in 3-person rings than in 6-person rings. More is sent with fixed matching than with random re-matching, with sequential decisions leading to less sent than with simultaneous decisions and random re-matching, but with little difference with fixed matching. So this study indicates that indirect reciprocity manifests even without a reputation system, but that the details of the environment are crucial.

Engelmann and Fischbacher (2009) separate out "pure" indirect reciprocity from strategic reputation building, using the Nowak and Sigmund's (1998) game; here the cost of giving is six, while the benefit is 15; the image score relates only to the past five times the recipient had acted as a donor. The main design innovation is that each participant had a public image score for either the first 40 periods or the last 40 periods of the 80 periods in a session; in the other periods one only had a private image score. This allows one to look at the difference in an individual's behavior depending on whether or not this behavior affects the information that others receive about her; when there is no public score, any indirect reciprocity is considered to be pure. They do find evidence of pure indirect reciprocity (perhaps driven by a taste for social efficiency), but donors with public scores have substantially higher giving rates than donors with private scores. Thus, reputation matters.

A number of other studies do not investigate indirect reciprocity, but do show that reputation systems for trustworthiness are effective. Bohnet and Huck (2004) consider the relative effectiveness of fixed matching and random re-matching when there is a history of return. In the first 10 periods of their design, their three treatments include random re-matching without information, fixed partnership, and random re-matching with information on the responder's history of return; in the second 10 periods, there is random matching without information in all three treatments. The trustworthy choice is most frequently observed in the fixed-partnership treatment, slightly less frequently in the random re-matching treatment with information about the responder's history, and much less frequently without any history provided. They find significant history effects, in that responders exhibit more trustworthiness after having been in a fixed partnership.

Bolton et al. (2004) provide a laboratory test of an online feedback mechanism modeled after those on Amazon, eBay, and Yahoo. The buyer chooses whether to buy an item and, if there is a purchase, the seller then decides whether to ship the item. If there is no purchase, each party keeps the initial endowment. In three 30-period treatments, while a feedback system leads to better social outcomes than would otherwise occur, it is less effective than a strategic repeated-game environment (with fixed pairings). There are substantial and significant differences in trust and trustworthiness across each pair of treatments. Thus, while a feedback system leads to better social outcomes than would otherwise occur, it is less effective than a strategic repeated-game environment.

---

[12] Gong and Yang (2010) find significant role for both standing and judging in a Prisoner's Dilemma experiment with second-order information for up to 10 periods.

[13] This is called "upstream" indirect reciprocity in Nowak and Sigmund (2005), in contrast to the "downstream" reciprocity we consider in this paper. Dufwenberg et al. (2001) also find upstream indirect reciprocity in a criss-cross version of a trust game, in which responders "return" to another person who is not the original first mover. Nowak and Sigmund (2005, p. 1292) distinguish between the two forms of indirect reciprocity as follows: " 'Upstream' reciprocity is based on a recent positive experience. A person who has been at the receiving end of a donation may feel motivated to donate in return. Individual B, who has just received help from A, goes on to help C. 'Downstream' reciprocity is built on reputation. Individual A has helped individual B and therefore receives help from C."
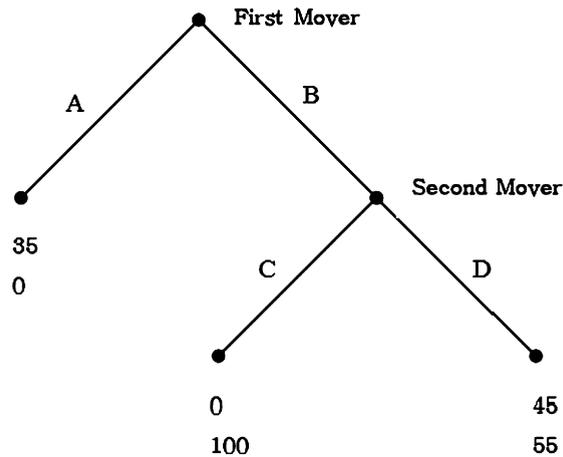
Fig. 1. The Game.

Keser (2004) has three treatments involving ratings (positive, negative, or neutral) and the investment game. The Baseline treatment involves no ratings, while the first mover observes only the very recent ratings about the responder in the short-run reputation treatment, and the long-run reputation treatment provides the distribution of *all* past ratings. She finds that information provision improves efficiency, with substantially more relative return with even a short-run reputation and a somewhat higher rate of relative return with a long-run reputation. Resnick et al. (2006) conducted a field experiment with an online reputation mechanism, in which an established eBay dealer made sales under both his regular identity and new seller identities; as predicted, the established identity fared better.

Duffy et al. (2008) use the trust game with a continuation probability of 80% after each period. Six people were in each group, with fixed roles and random re-matching every period. Their stage game has the same parameters as in our design, although the topics we examine are quite different.[14] Note that in their game (our ours), the design is asymmetric, with all of the efficiency gains resulting from the choice to trust.[15] The design varied whether first movers knew only their own history of play or were told the last decision of the current responder as a previous responder. They find considerably more cooperative behavior in the latter case, finding that a social norm of trust and reciprocity is difficult to sustain without providing reputational information.

3. Experimental design

We conducted our experiments at the Experimental Economics Laboratory, Shanghai University of Finance and Economics (SHUFE).[16] The participants were recruited from a campus-wide list of undergraduate students who had previously responded to advertisements in public courses or on the web. None of the participants had any experience with investment-game, indirect-reciprocity, or reputation-mechanism experiments. There were 16 matching groups (160 participants) in total, and no participant was permitted to participate in more than one session.

All laboratory sessions were computerized using Visual Basic 6.0; both the instructions and the information shown on the computer screen were in Chinese; a screenshot is provided in Fig. 7 in Appendix A. The participants played the game for 36 periods. In each period, each person had a 50% probability of being a first mover; we arranged the draws such that each person was a first mover 18 times and a second mover 18 times.[17] In each period, one was randomly and anonymously matched with one of the other nine participants. Sample instructions are provided in Appendix A.

The game structure (taken from Duffy et al., 2008) is shown in Fig. 1.

There were two sessions (four matching groups) in the Baseline treatment, where no information was given to the first mover before his decision; in each, there were two separate matching groups (i.e., 20 subjects played the game in the same room at the same time. They knew they were divided into two different groups and they knew they only interacted with people in the same group). The first mover either clicked the button 'A' or the button 'B' on her computer screen. If she clicked 'A,' the game was finished. Otherwise the game continued and the second mover clicked either 'C' or 'D.' The two sessions were conducted on the same day and took about 50 minutes each (including the time for reading instructions).

---

[14] They developed this parameterization so that in the absence of information, full trust and reciprocity can be sustained as an equilibrium under random matching and infinite horizon via a grim/contagious strategy. Lee and Xie (2007) develop a model in which the parameters in the game with information lead to a sequential equilibrium with both investing and returning. We use the parameters in their design, but look at different behavior and issues.

[15] Of course, this is not an inevitable feature (for example, Brandts and Charness, 2004 use a symmetric payoff design and permit efficiency gains in both directions), but seems to be quite common in this literature.

[16] Buchan and Croson (2004) find that Chinese students are more trusting and trustworthy than U.S. students.

[17] While they did not know that they would be in each role 18 times, they were aware that this was the expected number of times in each role.

Table 1
Trust and trustworthiness rates, by treatment.

Treatment	Matching groups	Trust rate	Trustworthiness rate	Rate of both combined, by pair
Baseline	4	0.242 [0.016] (720)	0.236 [0.032] (174)	0.057 [0.009] (720)
History of Return	6	0.541 [0.015] (1080)	0.656 [0.020] (584)	0.355 [0.015] (1080)
History of Trust	6	0.541 [0.015] (1080)	0.284 [0.019] (584)	0.154 [0.011] (1080)

Standard errors are in brackets. The number of observations in each cell is in parentheses.

The average payment was 37.8 yuan in RMB (100 points in Fig. 1 were equivalent to 3 yuan and the exchange rate was $1 = 6.85$ yuan), including a 10-yuan show-up fee. Since the average hourly wage in Shanghai for a college graduate is about 15–20 yuan, 37.8 yuan is a considerable amount for undergraduate students.

There were three sessions (six matching-groups) in each of the History of Return (henceforth, "HR") and the History of Trust (henceforth, "HT") treatments. In each case, the first mover could (beginning in the second period) click either or both of the "Summary Information" and "Detailed Information" buttons before choosing A or B. In the HR treatment, Summary Information provides the aggregate history of the second mover's behavior as a second mover and Detailed Information provides period-by-period history for the responder's behavior as a second mover; in the HT treatment, Summary Information provides the aggregate history of the second mover's behavior as a *first mover* and Detailed Information provides period-by-period history of the responder's behavior as a first mover.[18] In each treatment, all three sessions were conducted on the same day, each with two separate matching groups; these sessions took about 60 minutes. The average payment in both treatments was 48.35 yuan in RMB, including the 10-yuan show-up fee.

A design consideration is that we wished to examine the effect of a weaker form of information than the responder's history as a second mover, as it is less obvious that providing information about one's trusting behavior as a first mover should lead to either an efficient outcome B or the direct reciprocal action D.[19] Under HR, it is a rather clear strategy to invest in having a reputation as a trustworthy second mover, presuming that the first mover chooses to obtain historical information. Under HT, due to conceivable correlation between one's trusting and returning inclinations,[20] information on the responder's past trusting history would work as a very raw proxy for trustworthiness, at least in the beginning. But this proxy effect is not expected to sustain trust in community over the long haul, or even just beyond the initial periods. Thus, it is a challenge whether we could find sufficient evidence for indirect reciprocity, both hand-in-hand with any potential HT-as-proxy effect and eventually replacing it as the single-most driving motivation for trust. We discuss this issue in much greater detail in Section 5.

4. Results

In this section, we first present summaries of our data and statistical tests, then proceed to formal regressions to investigate the determinants of the observed behavior. As we shall see, the level of trust is significantly and substantially higher when a history of either trust or return is available, while the level of trustworthiness is much higher when a history of return is provided, but is not substantially different from the baseline when a history of trust is provided.

4.1. Data summary and statistical tests

Table 1 shows the average rates of B (trust) and D (trustworthiness) in each treatment. Providing either a history of return or a history of trust leads to an increase of the average trust rate of 30 percentage points, more than doubling the rate in the baseline. The trustworthiness rate is 40 percentage points higher than the baseline in the HR treatment, but is only five percentage points higher than the baseline in the HT treatment.

The most conservative statistical tests treat each matching group as only one independent observation (the rates for each of the 16 separate matching groups are shown in Tables B.1–B.3 in Appendix B). We use the Wilcoxon–Mann–Whitney rank sum test (see Siegel and Castellan, 1988) to compare the behavior across treatments. Trust rates are higher for all HR matching groups than for any Baseline matching groups; this leads to a test statistic of $Z = 2.59$, which indicates that the

---

[18] In each treatment, the summary information presented mentioned the number of times the individual was in the relevant role and how many times a particular action was chosen.

[19] Second, we chose not to use a stochastic ending both in order to observe the extent to which cooperative behavior decays (providing a rough measure of how much of the previous level of pro-social action is strategic) as a session progresses and because there is something of a loss of experimenter control over participants' beliefs when the ending period is unknown. Here we echo the argument from Bolton et al. (2004, p. 1591) that "participants would still guess at the market end [, setting up a confound] with other factors." We also note that the behavior in Seinen and Schram (2006) differed dramatically in the first (and certain) 90 periods and in those periods where the end was uncertain.

[20] For example, Altmann et al. (2008) find that first-mover and responder behavior in trust games is highly correlated. In Brandts and Charness (2000) and Charness and Rabin (2002), people also played in both roles, roughly corresponding to the first and second mover here; they find a fair degree of this type of 'consistency' for individuals across roles.
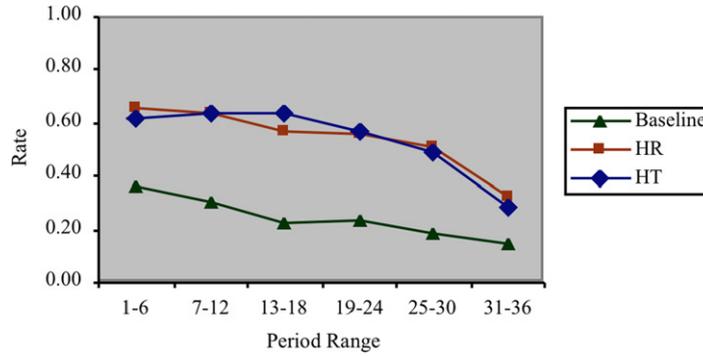
Fig. 2. Trust rates over time.

difference is significant at $p = 0.010$.[21] Since trust rates are also higher for every HT matching group than for any Baseline matching group, we again have a test statistic of $Z = 2.59$, indicating that the difference is significant at $p = 0.010$. Since the trust rates turn out to be identical in the HR and HT treatments, it is not surprising that the ranksum test shows no significant difference ($Z = 0.16$, $p = 0.873$).

It is also the case that trustworthiness rates are higher for all HR matching groups than for any Baseline matching groups; again this leads to a test statistic of $Z = 2.59$, with the difference being significant at $p = 0.010$. On the other hand, this test shows no significant difference in trustworthiness rates between the Baseline and HT treatments, with $Z = 0.64$, $p = 0.640$. And since trustworthiness rates are higher in every one of the HR matching groups than in any one of the HT matching groups, we have a test statistic of $Z = 3.06$, significant at $p = 0.002$.

Finally, the rate for which we observe that the first mover in a pair chose B and the second mover in the pair chose D is quite low in the Baseline, over one-third in HR, and intermediate in HT. This rate is higher in all HT sessions than in any Baseline session and is higher in nearly all HR sessions than in any HT session (one HT session had a higher rate than did one HR session). The corresponding test statistics are $Z = 2.59$ and $Z = 2.85$, significant at $p = 0.010$ and $p = 0.004$, respectively. The difference between the rates for HR and HT is clearly driven by the difference in the rates of trustworthiness across the HR and HT treatments. In our set-up, only the trust rate matters for social efficiency; however, if one is concerned with simultaneously achieving trust and trustworthiness in the pairs, clearly the HR treatment does best. Nevertheless, the HT treatment is a significant improvement over the Baseline treatment.

Since we have observations for each individual as both a first mover and a responder, we can examine on an individual level whether there is any correlation between one's trust and trustworthiness rates. The Spearman rank-correlation test gives a significant coefficient in the Baseline and HT treatments, with no significant correlation in the HR treatment. Furthermore, while there is no relationship between the number of times one is trusted and the rate of trustworthiness in the Baseline ($\rho = 0.006$, $p = 0.971$), we see a strong relationship between these in the HR treatment with a very high coefficient of $\rho = 0.744$ ($p = 0.000$). Indeed, this is how the direct reciprocity reputation system is supposed to work, since one would expect that people will tend to be more trusting of people who have been trustworthy.

For HT, there is a smaller but still significant coefficient of $\rho = 0.290$ ($p = 0.025$). However, the pattern differs greatly across the two halves of the HT treatment. Neither the correlation between trust and trustworthiness nor the correlation between trustworthiness and being trusted is significant in the latter half ($\rho = 0.182$, $p = 0.164$ and $\rho = 0.081$, $p = 0.538$ for respective correlations);[22] in comparison, we observe strong correlation between one's trust rate and one's rate of being trusted, even in the second half ($\rho = 0.6143$, $p = 0.000$).[23]

We next consider the patterns in behavior over time; recall that the environment is not stationary, since there is a known end to the session. This offers some insight regarding the degree to which behavior is strategic and the degree to which it reflects underlying social preferences. Figs. 2 and 3 show these patterns for trust and trustworthiness rates, respectively, while Fig. 4 shows the pattern for the combination of trust and trustworthiness for pairs.

We see a substantial decline in rates over time in all treatments for both trust and trustworthiness, and the rate at which a pair combined both of these also drops substantially. The pattern in trust rates over time is nearly identical (including the end-game rate of decline in trust rates) for the HR and HT treatments, while the decline in trust rates is a bit more moderate in the Baseline treatment. With respect to trustworthiness rates, there is a drop in the Baseline treatment in the early periods. The trustworthiness rate in the HR treatment decreases primarily in the ending periods, while the trustworthiness rate declines most dramatically in the first half HT treatment.[24] Finally, the rate at which pairs chose both trust and

---

[21] All statistical tests are two-tailed, except where otherwise indicated. All probabilities are rounded to three decimal places.

[22] In contrast, the correlation is highly significant ($p = 0.000$) in both halves of the HR treatment.

[23] This in fact corresponds to the data shown in Fig. 5 below and is evidence that people tend to trust other people who have a reputation for having trusted.

[24] In fact, D was rarely chosen in the last six periods of the HT treatment. Fig. 3 shows that the rate of trustworthiness even in the last six periods of the HR treatment (41.4%) is higher than the corresponding rate for the first six periods of the Baseline treatment (32.6%), which should represent some
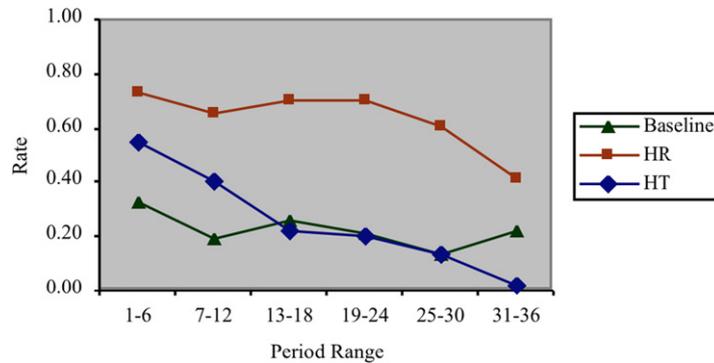
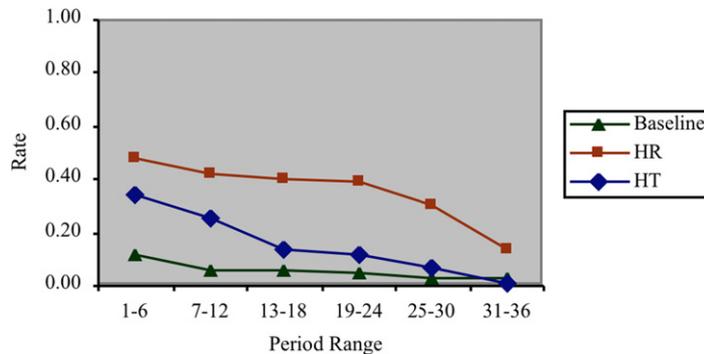**Fig. 3.** Trustworthiness rates over time.

Fig. 4. Rate of both trust and trustworthiness for pairs over time.

trustworthiness drops steadily in both the HR and HT treatments, while this rate is always low in the Baseline, particularly after the first few periods.

How responsive are first movers to historical information? First movers elected to check this history about 70% of the time in both of the HR and HT treatments, with no trend over time. Fig. 5 shows trust rates for historical information in the HR and HT treatments.[25]

There is a clear positive relationship between trust rates and historical information. When the responder's historical rate of trustworthiness or trust is no greater than 40%, the first mover chooses to trust only 13.4% or 22.8% of the time, respectively. By contrast, the first mover trusts 94.6% or 77.7% of the time when the historical rate of trustworthiness or trust, respectively, is greater than 90%. As a comparison, first movers who did not view the available history chose to trust 53.0% (50.8%) of the time in the HR (HT) treatment. There was much greater variance in the trust decision amongst those first movers who rarely viewed the available history, as it seems that such people are more inclined to the extremes.[26]

The response to reputation in HT seems flatter than that of HR, particularly from 0.401 on. This however is consistent with the manner in which the indirect reciprocity system in HT works. Rational reputation builders here are less inclined to contribute to discriminating amongst the second movers, which resembles a sort of secondary public-goods problem. Once their reputation score reaches a rather high level (so relatively safe from potential discrimination), they might think they could afford to withhold trust when this would only slightly reduce their score, as this avoids the second mover's (almost) sure no-return response.

A final consideration is whether it pays to trust the responder. From the standpoint of one's trust decision alone, since the rates of trustworthiness are lower than 7/9 in all treatments, one might conclude that trust does not pay.[27,28] However,

---

intrinsic inclination of direct reciprocity in the population. The relatively high trustworthiness rate at the end of the HR treatment may represent inertia, which has reinforcing effects.

[25]　It turns out that as long as some information is viewed there is no difference in behavior whether this is detail or summary information or both. Thus, we pool all observations where the first mover chose to observe the history.

[26]　The standard deviation for the number of times trust was chosen in HR was 3.65 when the first mover viewed more than four times (of 18), compared to 5.44 otherwise; the same comparison in HT gives 3.54 versus 6.62.

[27]　However, note that at the time of decision, the first mover observed a trustworthiness rate of at least as high as 7/9 nearly one-third (32.7%) of the time in HR.

[28]　Note that the issue in HR is very different than in HT. In the former, the first mover could see and use the whole return history and might weight it differently than in a linear form. So, one's decision about trust is a combination of both the degree of trustworthiness one has experienced in the past and the second mover's observed trustworthiness rate.
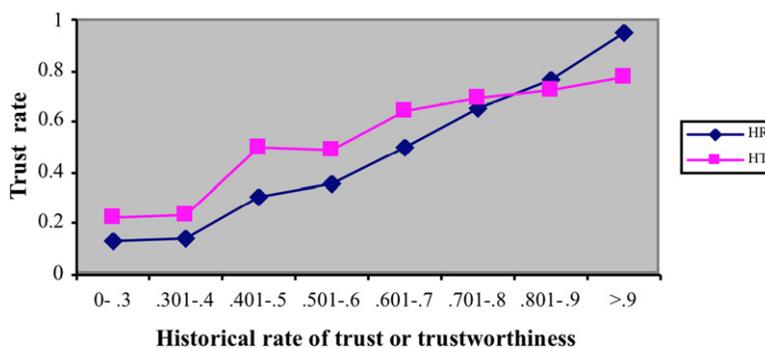
Fig. 5. Responder history and trust rates.

when we consider one's overall earnings (as both a first and second mover), it might well be advantageous to trust. In fact, there is a positive relationship between an individual's trust rate and her average payoff, although this is significant only in the HT treatment. The Spearman test gives $\rho = 0.130$ ($p = 0.324$) in the HR treatment and $\rho = 0.451$ ($p = 0.003$) in the HT treatment.[29] It appears to be worthwhile to establish a reputation for trusting when this is the history that is available for viewing.

4.2. Regression analysis

We perform probit regressions (with clustering on the matching-group level) for the determinants of trust and trustworthiness, taking into account variables that could be expected to affect such behavior. The regressions for the determinants of trust are shown in Table 2; we consider the behavior over all periods and also separately examine behavior in the first half and the second half of the game, as tests of structural consistency.

Note that the rate of trust decreases significantly over time in both the HR and HT treatments.[30] One category of determinants is related to the information variable. As already seen in Fig. 5, the revealed responder history in the overall data significantly affects the first mover's trust decision, laying down the foundation for the reputation system to work by providing an incentive for agents to build up reputation for later material reward while (potentially) sacrificing some of one's immediate payoff.[31]

The dummy View compensates for those situations when the first mover elected not to view the available information, in that the level of its coefficient compares viewing a (very poor) zero reputations with not viewing at all. The dummy D_View accounts for the case that no relevant information of the responder is available after electing to view, which occurs in the early periods of the HR treatment when the responder had not yet, at the time of this matching, previously had the opportunity to respond to being trusted.[32]

A second category of determinants concerns the personal experiences of the first mover. In both treatments, the more trustworthiness one has experienced (Exp_D), the more likely it is that one trusts, taking into account the (potentially) observed reputation of the responder. Conspicuously, however, this effect vanishes in the second half of HT. In addition, the first mover's most recent experience of trustworthiness has a particularly strong role. On the other hand, the first mover's last experience as a second mover has no significant effect. Note that the effect of experience of being trusted as a second mover (Exp_B) on one's trusting behavior changes from no significant effect in the first half, to significant in the second. We also note that one's own past behavior as a responder is significantly related to one's decision whether to trust in the

[29] A referee suggests that the insignificant positive correlation in HR between overall payoff and the trust rate could potentially be an artifact of the positive correlation between trust and trustworthiness, with the latter being the critical element and an expectation that the trust rate would then have a negative impact. As per his or her suggestion, we regress total payoff on both the trust rate and the trustworthiness rate. However, while the coefficient on trustworthiness is indeed positive and highly significant ($t = 3.57$), the coefficient on the trust rate is very slightly positive and insignificant ($t = 0.05$). Nevertheless, trusting does not immediately pay in HR. Thus, the weak positive correlation between overall payoff and trust rate could potentially be an artifact of the positive correlation between trust and trustworthiness and it is the latter that pays.

[30] A separate regression (not shown) for the Baseline treatment also shows a highly-significant decline in trust over time, with a coefficient of -0.023 for the period variable.

[31] Note that the coefficient for SM_Past_D is higher than the coefficient for SM_Past_B, perhaps implying more discrimination in the HR treatment; this would be consistent with the patterns observed in Fig. 5. However, in HR the FM only observes SM_Past_D and in HT the FM only observes SM_Past_B, so the two variables cannot be put in the same regression at the same time; thus, we see no clear way to test whether the difference is significant.

[32] There is no D_View in the HT regressions, since all responders in the first period become first movers in the second period and all first movers in the first period become SM in the second period. Beginning in the second period, first movers can click to view the responder's history of trust. Since first movers always have a move, a history is always available. This differs for second movers in HR, since the probability of a B choice is low enough that a second mover could experience only A moves for a considerable period of time.

Table 2
Determinants of trust.

Independent variables	HR (All) (1)	HR (1–18) (2)	HR (19–36) (3)	HT (All) (4)	HT (1–18) (5)	HT (19–36) (6)
Constant	0.111 (0.134)	0.147 (0.175)	0.490 (0.462)	0.196 (0.128)	−0.051 (0.152)	−0.238 (0.412)
Period	−0.040*** (0.005)	−0.053*** (0.016)	−0.056*** (0.013)	−0.024*** (0.004)	−0.018 (0.013)	−0.041*** (0.012)
View	−2.398*** (0.197)	−2.172*** (0.282)	−2.621*** (0.293)	−1.170*** (0.155)	−1.203*** (0.205)	−1.028*** (0.242)
Exp_B	−0.500** (0.243)	−0.530* (0.309)	0.132 (0.444)	−0.316 (0.202)	−0.003 (0.246)	1.186*** (0.386)
Exp_D	1.221*** (0.187)	1.381*** (0.290)	1.271*** (0.322)	0.476** (0.198)	0.517* (0.272)	0.379 (0.344)
Exp_B_Last	−0.081 (0.108)	−0.147 (0.164)	−0.087 (0.150)	0.181* (0.095)	0.104 (0.148)	0.183 (0.130)
Exp_D_Last	0.225* (0.116)	0.047 (0.213)	0.198 (0.146)	0.548*** (0.142)	0.469** (0.207)	0.643*** (0.202)
SM_Past_B	–	–	–	2.061*** (0.206)	1.971*** (0.266)	1.955*** (0.330)
SM_Past_D	3.592*** (0.255)	3.527*** (0.368)	3.804*** (0.381)	–	–	–
D_View	2.228*** (0.290)	2.077*** (0.356)	–	–	–	–
Action_D	0.596*** (0.201)	0.687*** (0.245)	0.233 (0.381)	0.510*** (0.154)	0.625*** (0.189)	0.347 (0.275)
N	1080	540	540	1080	540	540
LL	−529.24	−253.54	−271.32	−612.30	−300.20	−305.21
Wald χ^2	282.21	125.44	130.97	207.93	87.42	110.86
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

* indicates significance at $p = 0.10$ (two-tailed tests).
** indicate significance at $p = 0.05$ (two-tailed tests).
*** indicate significance at $p = 0.01$ (two-tailed tests).

Standard errors are in parentheses, with clustering by matching group. (All), (1–18), and (19–36) refer to the periods included in the regression. Exp_B (Exp_D) indicates the rate that the first mover has experienced trust (trustworthiness) as a responder (first mover). Exp_B_Last (Exp_D_Last) is the first mover's last experience of trust (trustworthiness) or non-trust (trustworthiness). SM_Past_B (SM_Past_D) indicates the rate at which the second mover has chosen to trust (be trustworthy), when viewed. D_View = 1 if there is no history available when viewed and is 0 otherwise (this variable is dropped due to collinearity in specification (3)). Action_D is the proportion of the time that the first mover has been trustworthy.

HT treatment, suggesting that there may be some false consensus effect at work.[33] However, there is no such relationship in the HR treatment, as seen in the lack of correlation between an individual's trust and trustworthy behavior noted earlier for this treatment.

Note that the coefficient for the proportion of the time that the first mover has been trustworthy (Action_D) is significant in the first half of both the HR and HT sessions, but not in the second half. This seems to imply that an initial intrinsic connection between trust and trustworthiness breaks down considerably in the session, in light of the *strategic* behavior of trusting fully based on information. In contrast, the main reputational variable – the second mover's trustworthiness rate (SM_Past_D) in the HR regressions and the second mover's trust rate (SM_Past_B) in the HT regressions – has a large and consistent impact.

The regressions for the determinants of return are shown in Table 3. We see that there is a decline in trustworthiness over time (coefficient of −0.051) in the HT treatment,[34] with a more moderate decline in HR (coefficient of −0.033), which reflects the trends displayed in Fig. 3. Aside from a time trend, the only relevant category of a determinant for trustworthiness is one's experience.[35] Here different aspects of one's experience are significant in the two treatments. In the HR treatment, the responder's experience with being trusted matters; however, the responder's experience with trustworthiness as a first mover does not significantly affect behavior. In the HT treatment, the responder's experience with being trusted does not matter, but the responder's experience with the trustworthiness of others has a significant effect. The responder's most recent experience with trusting or trustworthiness has a significant effect in the HR treatment, but very little effect in the HT treatment.[36]

---

[33] An informal definition of the false consensus effect (noted by Ross et al., 1977) is that people overestimate the likelihood that others choose as they do; here such an effect would imply that trustworthy second movers expect a higher trustworthiness rate and thus, all else equal, should be more likely to trust. See Engelmann and Strobel (2000) for further discussion.

[34] A separate regression (not shown) for the Baseline treatment also shows a similar trend with a coefficient of −0.064 for the period.

[35] Note that the coefficient on Action_B is insignificant, indicating that one's return decision as a second-mover is not correlated with one's trust decisions. Were this coefficient significantly positive, this would provide a hint that trust may have served as a proxy for trustworthiness.

[36] It is worth clarifying how to jointly interpret the coefficients on Exp_B_Max and Exp_B. Exp_B_Max is like a proxy for the subject's basic inclination to trustworthiness, as it reflects one's general rate of return as long as the HR reputation system works. One's on-the-spot return decision however is

Table 3
Determinants of return.

Independent variables	HR (All) (1)	HR (1–18) (2)	HR (19–36) (3)	HT (All) (4)	HT (1–18) (5)	HT (19–36) (6)
Constant	−0.648***	−0.908***	0.966	−0.821**	−0.824**	−5.507
	(0.199)	(0.259)	(0.642)	(0.317)	(0.350)	(N/A)
Period	−0.033***	−0.029*	−0.079***	−0.051***	−0.078***	−0.046*
	(0.006)	(0.017)	(0.020)	(0.008)	(0.020)	(0.026)
Exp_B	−1.329***	−1.050***	−3.700***	−0.304	−0.474	1.215
	(0.316)	(0.366)	(1.253)	(0.391)	(0.421)	(1.551)
Exp_D	0.195	0.159	−0.114	0.602***	0.496**	1.651**
	(0.301)	(0.422)	(0.495)	(0.210)	(0.226)	(0.747)
Exp_B_Max	0.200***	0.225***	0.296***	0.069**	0.083***	−0.023
	(0.023)	(0.028)	(0.078)	(0.028)	(0.031)	(0.100)
Exp_B_Last	0.340**	0.209	0.461**	0.246	0.473	4.743***
	(0.153)	(0.232)	(0.216)	(0.337)	(0.388)	(0.958)
Exp_D_Last	0.323**	0.469*	0.159	0.076	0.067	−0.158
	(0.156)	(0.256)	(0.211)	(0.139)	(0.175)	(0.269)
Action_B	−0.058	−0.359	0.751	−0.033	−0.053	−0.459
	(0.270)	(0.329)	(0.501)	(0.247)	(0.261)	(0.896)
N	584	335	249	584	342	242
LL	-306.46	−159.42	−139.44	−297.56	−203.88	−89.81
Wald χ^2	113.16	71.01	48.45	89.58	44.15	741.02
p-value	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

* indicates significance at $p = 0.10$ (two-tailed tests).
** indicate significance at $p = 0.05$ (two-tailed tests).
*** indicate significance at $p = 0.01$ (two-tailed tests).
Standard errors are in parentheses, with clustering by matching group. (All), (1–18), and (19–36) refer to the periods included in the regression. Exp_B (Exp_D) indicates the rate that the responder has experienced trust (trustworthiness) as a responder (first mover). Exp_B_Max is the number of times in a session that the responder is trusted. Exp_B_Last (Exp_D_Last) is the responder's last experience of trust (trustworthiness). Action_B is the rate at which the responder has trusted as a first mover.

Finally, we note a strong positive relationship in both treatments between the total number of times that an individual is trusted in a session and the likelihood that he chooses the trustworthy action, even when taking into consideration the other experience variables. It would seem that people who are given more opportunities to make a choice as a responder are more willing to be cooperative. In the HR treatment it is also the case that people with a better history of trustworthiness will be trusted more frequently, so here the direction of causality is unclear. Nevertheless, there is no such link in the HT treatment.

Comparing periods 1–18 with periods 19–36, there are few significant differences in Table 3. We do see that the responder's most recent experience with trust (Exp_B_Last) is not significant in the first half of the HR and HT games, but becomes significant in the second half of these games; in the latter case, this crowds out the significance of the number of times that the responder has been trusted (Exp_B_Max).[37]

5. Discussion

We have seen that reputation systems are effective in inducing trust, at least until near the end of the game. While in a game without a definite end we might expect this pattern to continue indefinitely, the terminal rates give some indication of the residual levels of trust after reputation is no longer a factor. Since we see rates appreciably above zero even at the end of the sessions, there is some scope for non-strategic cooperation, driven either by considerations of reciprocity or a desire to increase the total social payoff. More interestingly, while previous experimental work has shown the value of encouraging a reputation for having been trustworthy (building a reputation based on direct reciprocity), we find that trust levels are just as high with a reputation system for being trusting.[38] Both factors play a statistically significant role, as documented in our regression analysis.[39]

---

positively affected by one's *last period* experience of being trusted or meeting a trustworthy partner (Exp_B_Last and Exp_D_Last). In general, however, one has a slightly higher incentive to build reputation when it is low, which is reflected in a negative coefficient for Exp_B via correlation.

[37] While the coefficient of 4.743 may seem high, note that it is offset by the coefficient for the constant term changing to −5.507 within this regression. Also note that the total level of return is very low in 19–36.

[38] The fact that the levels of trust are the same in the HR and HT treatments is indeed surprising; nevertheless, some caveats may apply. Note that the trust history has a direct effect on trust (though based on indirect reciprocity), whereas the return history can only work indirectly. If there are first movers who do not feel that preferences are stable over time, this will weaken the impact history of return has on the trust rate, since for such first movers the signal of past behavior is irrelevant with respect to predicting future behavior; on the other hand, if trust is based on indirect reciprocity, then trust is a reward for past behavior and not an anticipation of future behavior. In addition, the HT treatment may also do quite well since the efficiency gains are all on the first stage, so if a system of indirect reciprocity replaces one of reputation building for trustworthiness, this will lead to the same efficiency gains.

[39] Note the positive effect of Exp_D on trust (direct reciprocity) and return (upstream indirect reciprocity) decisions.
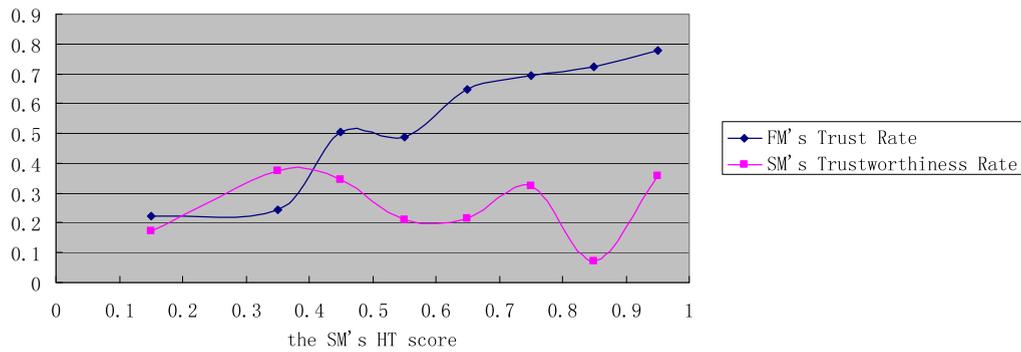
Fig. 6. First- and second-mover actions based on viewed HT score.

For any indirect reciprocity system to work, first movers in our HT treatment must collectively discriminate amongst the responders based on their trusting rates, else it would not pay to invest in a reputation. Fig. 5 tells us that this is true in our experiment. In fact, Fig. 2 reveals that HT generates significantly more trust than the baseline. In addition, while HT leads to just as much trust as did HR (with its direct-reciprocity reputation system), Fig. 3 shows the fundamental difference between the direct and indirect reciprocity network in this environment, namely, that the rate of return drops to very little (a rate similar to that in the Baseline) over time in HT.

One fundamental issue is whether the trusting behavior we observe in the HT treatment is really indirect reciprocity or if observed trust is simply a proxy for trustworthiness. The latter interpretation would seem quite reasonable, particularly since other studies have found a correlation between trust and trustworthiness. Such a proxy effect is not inconsistent with the analysis for the first half (periods 1–18) of the sessions. However, recall that while we find a highly-significant relationship between trust and trustworthiness in the first half, there is no significant relationship in the second half of the game.[40] Fig. 6 clearly illustrates how the discrepancy between the expected rate of trust and that of trustworthiness widens with the HT score once this score is at least 0.45.

Second, there is a strong effect (consistent over time) of one's experience with trustworthiness (Exp_D) on trust in the HR treatment, but only a modest one (vanishing over time) in the HT treatment; this goes against the notion that trusting decisions are based on a higher expected return rate coded in a higher HT score. Finally, the results show that during the first half of the HT experiments, trustworthiness (Action_D) has a significant impact on the trust decision, but it is no longer so during the second half; in contrast, the second-mover's HT score (SM_Past_B) has a very strong impact throughout. This evidence goes against the hypothesis that the HT score proxies for the trustworthiness. Thus, our results imply a substantial role for indirect reciprocity.

So, we feel that the HT score may code (imperfectly) for trustworthiness in the early periods, giving the first mover additional incentive to discriminate based on the perceived likelihood of receiving direct reciprocation in the current period, beyond own image-score building incentives. However, as time goes by and the indirect reciprocity system works on its own, the pure score-building motivation outweighs the secondary pull of expected trustworthiness, as the latter gradually phases out.

Within a functioning indirect-reciprocity reputation system, strategic reputation builders are subject to conflicting motivations along the process. First, if the discrimination within the system is very strong, rational players may prefer maximizing their own image score by simply trusting everyone and, hence, not participating in discrimination. However, this would give non-trusting types (and thus inefficiency) a chance to creep back into the population.[41] On the other hand, if discrimination is imperfect within the system, then rational players with an already high image score might have incentives not to "over-invest" and thus selectively withhold trust without being put in great danger of immediately being the target of discrimination. The issue becomes how such a set of conditional and sophisticated discrimination strategies can be put together to make the reputation system work. Associated with this, it is not a trivial question as to why the players shall discriminate if image score were their only concern of interest.

One answer to this is to treat social efficiency, as in Charness and Rabin (2002) and Engelmann and Strobel (2004), as a secondary motivation. Assume the players aim at maintaining an optimal level of image score below that of the maximum. They would allocate their own quota of no-trust to those matches with a low-score responder. This would help discrimination become common practice that enables the reputation system to work, for greater social benefit. Direct

---

[40] A correlation test gives $\rho = 0.418$ ($p = 0.000$) in the first 18 periods of HT, compared to $\rho = 0.081$ ($p = 0.538$) in the last 18 periods.

[41] Nowak and Sigmund (2005, p. 1293) state, "Two features of this model were immediately apparent. One is the paradoxical nature of the discriminating strategy. In terms of rational game theory, why should players care about the scores of others rather than just about their own payoff, and why should they decrease their own score (and thus their likelihood of receiving help on later occasions) by withholding help from low-scorers? Lower scores can cause lower payoffs. The second issue concerns the lack of stability of the cooperative outcome. The simulations display occasional bursts of defection, which are based on a previous build-up of undiscriminating altruists. In a population of discriminators, unconditional cooperators can increase by random drift and eventually invite the invasion of defectors (Fig. 3)."

personal comparison with the responder's scores may also induce discrimination in aggregate, via the psychological paths of introspection, extrapolation, or something closely related to "false consensus effect." If the responder has a higher score, the trustor might have good reason to believe that she would be punished in her own responder role by others for the very same reason she would be punishing her current would-be responder. Thus, she would be more inclined to be trusting here. However, if (by comparison) the opposite is the case, the first mover might conclude that her own current score is high enough to be safe from similar discrimination and thus be more inclined to strategically use this chance to secure expected higher payoff by withholding trust here. The total aggregate effect would be *collective discrimination*.

6. Conclusion

We test whether trust can be sustained in a strategic environment by two different forms of reputation systems. While we confirm the results of previous studies that providing information concerning the responder's past trustworthiness can sustain trust, our new result is that trust can also be sustained by providing information concerning the responder's past history of being trusting. In fact, the overall levels of trust with these two reputation systems are the same, with a much higher trust rate than when there is no reputation system in place.

What is also surprising is that this high rate of trust is sustained without any meaningful level of trustworthiness along the way. One would expect that the willingness to trust to be adversely affected by a lack of positive reciprocal response. Yet people continue to trust in order to build a good reputation despite the poor immediate return on this investment, implicitly recognizing that many people will make trust choices based on the other person's history of trust. In fact, even though it does not directly pay to trust in our history of trust treatment,[42] the gain one receives from establishing a reputation for trusting outweighs this immediate cost, as we find a positive relationship between an individual's trust rate and her average payoff in the HT treatment.[43]

Our finding suggests that a broader range of reputation systems can be effective than has been documented in the literature. Note that List (2006) offers insightful field evidence that reputation building in the real market, which is reminiscent of the sequential prisoners' dilemma game, can be indeed highly sensitive to the environment. So, one cannot expect that any reputation would always work. More research is needed on clarifying the conditions for any particular reputation system's eventual success.

Acknowledgments

Appendix A. Sample instructions

Instructions (History of Trust)

Welcome to our experiment. You will earn lab money depending on how you and others decide. At the end of experiment, you can have your lab money exchanged for RMB. You will get 3 RMB for 100 points. (1 USD = ca. 7 RMB.)

There are two groups of 10 participants each in this room. The 10 participants within the same group will be randomly re-matched in pairs among themselves round-by-round for a total of 36 rounds. At the beginning of each round, you will be randomly selected to be either the first mover or the second mover, as displayed on your screen. The matching is anonymous. However, you will take the role of the first mover and second mover at least once each in every 4 rounds.

In each round, the first mover is to first choose between A and B. If A is chosen, by clicking the button "A" under "Your decision is" as shown on the screenshot, this round ends. The first mover gets 35 pts and the second mover gets 0 pt. If B is chosen, then it is the second mover's turn to choose between C and D. And she will be able to click the buttons "C" or "D" under "Your decision is." If C is chosen, the first mover gets 0 pt while the second mover gets 100 pts. If D is chosen, the first mover gets 45 pts while the second mover gets 55 pts.

Beginning in the second round, the first mover will be able to view some information about the matched second mover's past decisions *as first mover*. As shown in the screenshot below, he or she could click on either "Summary" or "Detail." With Summary, he would see the numbers of times his partner chose A and B in all past rounds as first mover herself. With Detail, he will get the exact sequence of those decisions with precise round information.

[**History of Return**: Change *as first move*r to *as second mover.* And change A/B to C/D subsequently.]

---

[42] The expected direct payoff from trusting is only 12.78, compared to 35 from not trusting.

[43] Normalizing both payoffs and trust rate with respect to the group means, the correlation becomes insignificant (the *t*-statistic on the coefficient of the difference in trust rates is 1.24, giving a *p*-value of 0.219). So, there is no significant correlation that undermines survival of trust without such information system, even within the closed group. On the global level, the significant positive correlation just mentioned in the text lays down the foundation for its further evolution into dominance, in the assortative-matching notions of evolution dynamics (see Sober and Wilson, 1998, and Bergstrom, 2002).
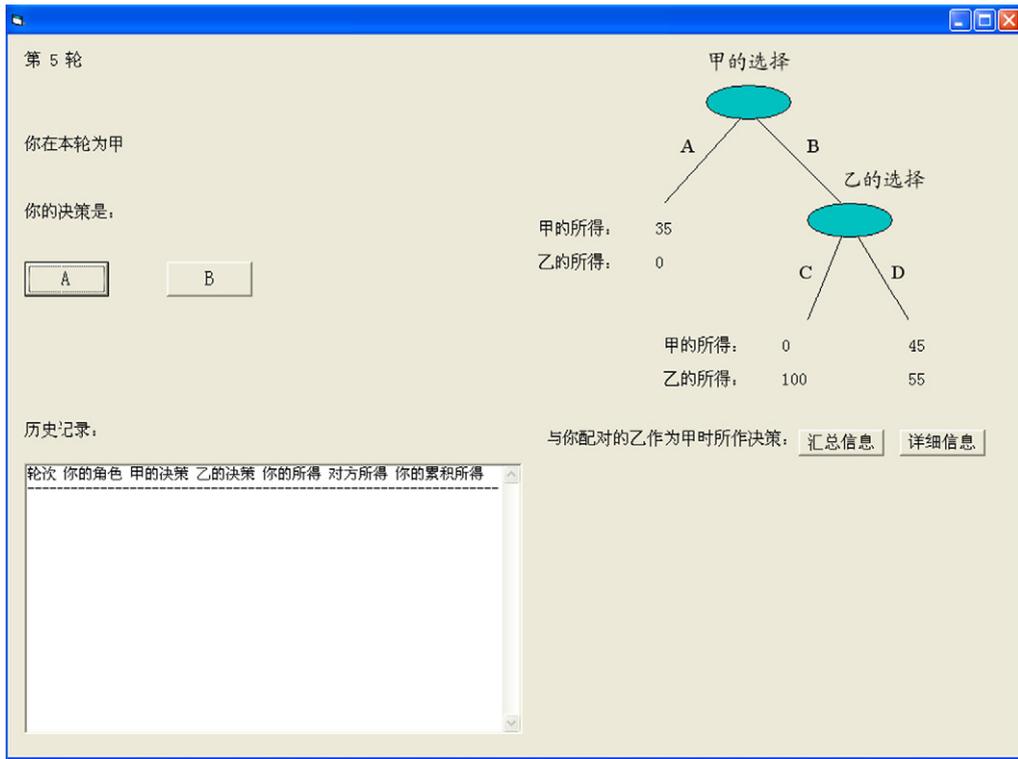
Fig. 7. Screenshot of instructions.

At the end of each round, the first-mover and second-mover decisions including the associated payoffs in the pairing will be recorded under "History" on the screen, as an add-on entry round-by-round. (See Fig. 7.)

Appendix B. Trust and trustworthiness rates by matching group

Table B.1
Trust and trustworthiness rates, Baseline sessions.

Matching group	Trust rate	Trustworthiness rate	Rate of both combined, by pair
1	0.233 [0.022]	0.071 [0.028]	0.017 [0.007]
	(360)	(84)	(360)
2	0.278 [0.024]	0.320 [0.047]	0.089 [0.015]
	(360)	(100)	(360)
3	0.333 [0.025]	0.283 [0.041]	0.094 [0.015]
	(360)	(120)	(360)
4	0.122 [0.017]	0.227 [0.064]	0.028 [0.009]
	(360)	(44)	(360)

Standard errors are in brackets. The number of observations in each cell is in parentheses.

Table B.2
Trust and trustworthiness rates, HR sessions.

Matching group	Trust rate	Trustworthiness rate	Rate of both combined, by pair
1	0.556 [0.026]	0.580 [0.035]	0.322 [0.025]
	(360)	(200)	(360)
2	0.489 [0.026]	0.659 [0.036]	0.322 [0.026]
	(360)	(176)	(360)
3	0.517 [0.026]	0.559 [0.037]	0.289 [0.024]
	(360)	(186)	(360)
4	0.750 [0.023]	0.778 [0.025]	0.584 [0.025]
	(360)	(270)	(360)
5	0.528 [0.026]	0.695 [0.033]	0.367 [0.026]
	(360)	(190)	(360)
6	0.406 [0.026]	0.603 [0.041]	0.245 [0.024]
	(360)	(146)	(360)

Standard errors are in brackets. The number of observations in each cell is in parentheses.

Table B.3
Trust and trustworthiness rates, HT sessions.

Matching group	Trust rate	Trustworthiness rate	Rate of both combined, by pair
1	0.611 [0.026]	0.336 [0.032]	0.205 [0.021]
	(360)	(220)	(360)
2	0.483 [0.026]	0.230 [0.032]	0.111 [0.017]
	(360)	(174)	(360)
3	0.356 [0.025]	0.281 [0.040]	0.100 [0.016]
	(360)	(128)	(360)
4	0.617 [0.026]	0.189 [0.026]	0.117 [0.017]
	(360)	(222)	(360)
5	0.694 [0.024]	0.376 [0.031]	0.261 [0.023]
	(360)	(250)	(360)
6	0.483 [0.026]	0.264 [0.034]	0.128 [0.018]
	(360)	(174)	(360)

Standard errors are in brackets. The number of observations in each cell is in parentheses.

References

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

Altmann, S., Dohmen, T., Wibral, M., 2008. Do the reciprocal trust less? Econ. Letters 99, 454–457.

Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games Econ. Behav. 10, 122–142.

Bergstrom, T., 2002. Evolution of social behavior: Individual and group selection. J. Econ. Perspect. 16, 67–88.

Bohnet, I., Huck, S., 2004. Repetition and reputation: Implications for trust and trustworthiness when institutions change. Amer. Econ. Rev. Papers Proceed. 94, 362–366.

Bolton, G., Katok, E., Ockenfels, A., 2004. How effective are electronic reputation mechanisms? An experimental investigation. Manage. Sci. 50, 1587–1602.

Bolton, G., Katok, E., Ockenfels, A., 2005. Cooperation among strangers with limited information about reputation. J. Public Econ. 89, 1457–1468.

Brandts, J., Charness, G., 2000. Hot vs. cold: Sequential responses in simple experimental games. Exper. Econ. 2, 227–238.

Brandts, J., Charness, G., 2004. Do labour market conditions affect gift exchange? Some experimental evidence. Econ. J. 114, 684–708.

Buchan, N., Croson, R., 2004. The boundaries of trust: Own and others' actions in the US and China. J. Econ. Behav. Organ. 55, 485–504.

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

Coleman, J., 1990. Foundations of Social Theory. Harvard University Press, Cambridge, MA.

Duffy, J., Lee, Y., Xie, H., 2008. Social norms, information, and trust among strangers: An experimental study. Mimeo.

Dufwenberg, M., Gneezy, U., Güth, W., Van Damme, E., 2001. Direct vs. indirect reciprocity: An experiment. Homo Oeconomicus 18, 19–30.

Engelmann, D., Fischbacher, U., 2009. Indirect reciprocity and strategic reputation building in an experimental helping game. Games Econ. Behav. 67, 399–407.

Engelmann, D., Strobel, M., 2000. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. Amer. Econ. Rev. 94, 857–869.

Engelmann, D., Strobel, M., 2004. The false consensus effect disappears if representative information and monetary incentives are given. Exper. Econ. 3, 241–260.

Fehr, E., 2004. Don't lose your reputation. Nature 432, 449–450.

Fukuyama, F., 1995. Trust: The Social Virtues and the Creation of Prosperity. Free Press, Glencoe, IL.

GartnerG2, 2002. Online transaction fraud and prevention get more sophisticated, http://www.gartnerg2.com/rpt/rpt-0102-0013.asp.

Gong, B., Yang, C., 2010. Reputation and cooperation: An experiment on prisoner's dilemma with second-order information. Mimeo, http://ssrn.com/abstract=1549605.

Greiner, B., Levati, V., 2005. Indirect reciprocity in cyclical networks: An experimental study. J. Econ. Psych. 26, 711–731.

Kahneman, D., Knetsch, J., Thaler, R., 1986. Fairness and the assumptions of economics. J. Bus. 59, S285–S300.

Keser, C., 2004. Trust and reputation building in e-commerce. Mimeo.

Lee, Y., Xie, H., 2007. Social norms and trust among strangers. Mimeo.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proc. R. Soc. Lond.: Biol. Sci. 268, 745–753.

List, J., 2006. The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. J. Polit. Econ. 114, 1–37.

Milinski, M., Semmann, D., Bakker, T., Krambeck, H., 2001. Cooperation through indirect reciprocity: Image scoring or standing strategy? Proc. R. Soc. Lond.: Biol. Sci. 268, 2495–2501.

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

Nowak, M., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.

Ohtsuki, H., Iwasa, Y., Nowak, M., 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457, 79–82.

Ostrom, E., Walker, J., 2002. Trust and Reciprocity: Interdisciplinary Lessons for Experimental Research. Russell Sage Foundation, New York.

Ostrom, E., Gardner, R., Walker, J., 1992. Covenants with and without a sword: Self-governance is possible. Amer. Polit. Sci. Rev. 86, 404–417.

Panchanathan, K., Boyd, R., 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432, 499–502.

Resnick, P., Zeckhauser, R., Swanson, J., Lockwood, K., 2006. The value of reputation on eBay: A controlled experiment. Exper. Econ. 9, 79–101.

Rockenbach, B., Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444, 718–723.

Ross, L., Greene, D., House, P., 1977. The false consensus effect: An egocentric bias in social perception and attribution processes. J. Exper. Social Psych. 13, 279–301.

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

Siegel, S., Castellan, N.J., 1988. Nonparametric Statistics for the Social Sciences. McGraw–Hill, Boston.

Sober, E., Wilson, D., 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

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

Yang, C., Weimann, J., Mitropoulos, A., 2006. An alternative approach to explaining bargaining behaviour in simple sequential games. Pacific Econ. Rev. 11, 201–221.