Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2015, Article ID 234528, 12 pages
http://dx.doi.org/10.1155/2015/234528

Research Article

Intermediaries in Trust: Indirect Reciprocity, Incentives, and Norms

Giangiacomo Bravo,[1] Flaminio Squazzoni,[2] and Károly Takács[3]

[1] *Department of Social Studies, Linnaeus University, Universitetsplatsen 1, 35195 Växjö, Sweden*
[2] *Department of Economics and Management, University of Brescia, Via San Faustino 74B, 25122 Brescia, Italy*
[3] *MTA TK Lendület Research Center for Educational and Network Studies (RECENS), Hungarian Academy of Sciences, Országház Utca 30, Budapest 1014, Hungary*

Correspondence should be addressed to Giangiacomo Bravo; giangiacomo.bravo@lnu.se

Received 2 September 2014; Accepted 4 December 2014

Academic Editor: Vincent Buskens

Any trust situation involves a certain amount of risk for trustors that trustees could abuse. In some cases, intermediaries exist who play a crucial role in the exchange by providing reputational information. To examine under what conditions intermediary opinion could have a positive impact on cooperation, we designed two experiments based on a modified version of the investment game where intermediaries rated the behaviour of trustees under various incentive schemes and different role structures. We found that intermediaries can increase trust if there is room for indirect reciprocity between the involved parties. We also found that the effect of monetary incentives and social norms cannot be clearly separable in these situations. If properly designed, monetary incentives for intermediaries can have a positive effect. On the one hand, when intermediary rewards are aligned with the trustor's interest, investments and returns tend to increase. On the other hand, fixed monetary incentives perform less than any other incentive schemes and endogenous social norms in ensuring trust and fairness. These findings should make us reconsider the mantra of incentivization of social and public conventional policy.

1. Introduction

A trust relationship is an exchange where at least two parties interact, that is, a trustor and a trustee, and in which there is a certain amount of risk for the former. If the trustor decides to place trust, the trustee can honour or abuse it. If honouring trust is costly, as what happened in one-shot interactions and sometimes even in repeated exchanges, the trustee will have no rational incentive to be trustworthy. Knowing this, the trustor is hardly likely to make the first move [1].

Understanding how trust can be established in such hostile situations is of paramount importance. One of the most interesting sociological explanations suggests that social and economic exchanges are embedded in social contexts where certain norms and roles have evolved to mediate between individual interests. For instance, intermediaries might act as advisories and mediators between the parties involved and reputation or gossip can help to spread information about unknown partners that helps trustors to take the risk of interaction [1–3].

Recent experimental results have shown that individuals can overcome distrust and cooperate more frequently if behaviour in the exchange is observed by a third party [4–8]. This happens even when the opinion of a third party has no consequence on the payoffs of the individuals and reputational building strategies are ruled out [9]. This indicates that, in many real situations, third parties can reduce information asymmetry and temptations of free riding, induce mutual trust, and ensure collective benefit. This requires understanding why and under which conditions information from third parties should be trusted by trustors and what type of incentives can make judgements or recommendations by third parties credible to the eyes of trustors. Indeed, first, intermediaries' opinion is often subjective. Secondly, it is formulated on trustee's past behaviour in situations where the trustees could have strategically mimicked trustwhorty signals in view of benefits from future trustors. This transforms the problem of trust in a "secondary problem of trust" [10]. Here, the challenge is to understand under what conditions, in anonymous exchanges with unknown partners, intermediaries could be considered reliable and how they should be motivated to provide information that increases trust.

For instance, let us consider certain important social and economic interactions, such as the trust relationships between employees and managers in big companies or between two potential partners for a relationship. In these cases, intermediaries are called on to express their opinion on certain attributes or behaviour, which are crucial to create trust. Sometimes they do so voluntarily, without receiving any material payoffs, such as someone recommending a friend as a partner to another friend. In other cases, intermediaries are financially motivated professionals, such as an HR manager recommending an employee to be upgraded to his/her company manager. Therefore, while in certain spheres of social life, the social function of intermediaries has been institutionalized through material incentives and roles, in other situations informal or voluntary norms have developed.

The aim of our paper was to examine these trust problems in triadic relations (i.e., between trustors, intermediaries, and trustees) to better understand conditions that could transform the intermediary opinion in a trust carrier. We conducted two experiments where subjects played a modified version of the repeated investment game with intermediaries added to the typical trustor/trustee dyadic relation. We manipulated incentives and roles of intermediaries to test their impact on cooperation in particularly hostile conditions. For this, we meant a situation where (i) intermediaries formulated their opinion on the trustee behaviour on a limited set of information and (ii) their opinion was not public and (iii) did not have long-term consequences on the material payoffs of the trustees. In this situation, intermediaries had only a limited set of information (i) and bad standing was not so risky for trustees (ii-iii). Hence, our experimental situation was intentionally designed so that intermediary opinion was poorly credible for trustors. We examined the importance of indirect reciprocity considerations and their interplay with material incentives and social norms. Furthermore, we tested different ways to align intermediary's incentives, for example, respectively, with the trustors' or the trustees' interests, and the impact of roles' rotation.

We compared results from two laboratory experiments Experiment 1, first reported in [11], and Experiment 2, previously unpublished. The latter was designed to carefully examine the effect of indirect reciprocity on the behaviour of intermediaries—and, more generally, on trust at the system level—since this mechanism was crucial in the first experiment. The rest of the paper introduces our research background and hypotheses (Section 2), presents the design of the two experiments (Section 3) and their outcomes (Section 4), and finally discusses certain social and policy implications of the results (Section 5).

2. Hypotheses

While dyadic embeddedness is important to explain trust and cooperation in situations involving stable relationships between two actors, in modern society trust is often mediated by agents who facilitate the exchange between trustors and trustees when they cannot rely on past experience [1, 12, 13]. Important examples of this have been empirically found in the development of trust between suppliers and customers in a variety of situations, such as in electronic markets on the web [14, 15] and in the US venture capital market [16]. In these cases, any chance for trust is undermined by information asymmetry between the involved parties and does not simply imply all-none choices, such as trusting or not, but requires a pondered rational calculus of trust investment.

Let us consider the *investment game*, a typical framework to model trust problems [17]. First, Player A (the trustor) receives an initial endowment of d_{A} points, with a fixed exchange rate in real money. Then, A is called to decide the amount I between 0 and d_{A} to send to Player B (the trustee), keeping the part $(d_{\mathrm{A}} - I)$. The amount sent by A is multiplied by $m > 1$ and added to B's own endowment d_{B}. (Note that in some investments games, including [17], B players had no endowment; that is, $d_{\mathrm{B}} = 0$. Trustor investments tend to be lower in experiments where $d_{\mathrm{B}} > 0$ [18].) Then, B is called to decide the share of the amount received (plus his/her endowment) to return to A. As before, the amount R returned by B can be between 0 and $(d_{\mathrm{B}}+mI)$. The amount not returned represents B's profit, while R is summed to the part kept by A to form his/her final profit. Therefore, the player's profit is as follows:

$$\begin{aligned} P_{\mathrm{A}} &= d_{\mathrm{A}} - I + R, \\ P_{\mathrm{B}} &= d_{\mathrm{B}} + mI - R. \end{aligned} \tag{1}$$

The structure of the game implies that the trustor can have an interest in investing on condition that the expected returns are higher than his/her own investments, that is, if $R = q(d_{\mathrm{B}} + mI) > I$, where q is the proportion returned by the trustee. However, as the trustor can rationally presume that the trustee has no interest in returning anything, it is rationally expected that the trustor will not invest and trust will not be placed, giving rise to a suboptimal collective outcome of $(d_{\mathrm{A}} + d_{\mathrm{B}})$ while the social optimum $(md_{\mathrm{A}} + d_{\mathrm{B}})$ could have been reached with sufficient high levels of trust. However, empirical evidence contradicts this prediction. A recent review of 162 experimental replications of the investment game indicated that, on average, trustors invested about 50% (range 22–89%) of their endowment and trustors returned 37% (range 11–81%) of their amount [18].

In this type of games, the crucial challenge is to understand the mechanisms through which the trustor estimates the trustworthiness of the trustee by using available information. Following Coleman [1], we can identify two possible information sources for trustors: (i) direct knowledge of past behaviour of the trustee and (ii) knowledge of trustee behaviour obtained by a third party with positions and interest differently aligned to those of the other parties involved. In both cases, information on past behaviour of the trustee may help the trustor in predicting the trustworthiness of the trustee, creating in turn reputational incentives for the trustee to overcome any temptation of cheating in view of future benefits.

While it is widely acknowledged that knowing the past behaviour of a trustee can increase a trustor's investment and cooperation in dyadically embedded interaction [19–23], the case of triadic relations is more interesting as these relationships can compensate for the lack of direct knowledge and contacts between actors, so enlarging the social circles and the extent of the exchange [24]. This requires us to understand the complex triangulation of the exchange and especially the role of trust intermediaries, who might have either analogous or different positions and interests to the trustors. When positions and interests between the trustor and the intermediary are aligned, the intermediary can act in the trustor's interest and the latter seriously considers the intermediary's opinion so that his/her decision will reflect the available reputational information. When the intermediary and the trustor do not have aligned interests, the outcome of the exchange is heavily dependent on the motivations behind the intermediary's action [1].

In order to represent triadic relations in trust situations, we modified the standard investment game framework by adding to the trustor (called Player A) and the trustee (called Player B) a third player, that is, the intermediary (called Player C). We assumed that the intermediary could observe B past behaviour (i.e., the amount of returns sent to A compared to the A investment) and was called to decide whether to provide A with honest and accurate information or not. If credible to the eye of A, information by C is expected to help A regulate his/her investment. If this is so, B Players have a rational interest in building a good standing at the eye of C and so behave more fairly with A. Therefore, if intermediaries are trusted both by the trustors and by the trustees it is expected to generate higher levels of cooperation and fairness.

When modelling intermediaries' behaviour, for the sake of simplicity we assumed that C could only choose between two levels of fairness: providing A with an accurate evaluation of B (=high fairness) or sharing inaccurate or deceptive information (=low fairness). Adapting Frey and Oberholzer-Gee's "motivation crowding-out" formalization [25], we also assumed that each intermediary chooses a level of fairness (f) that maximizes his/her expected net benefit as follows:

$$\max\left[p\left(f\right)(b+e)+d\left(e,f\right)-k\right]. \qquad (2)$$

Since $m > 1$, any increase of investments leads to higher stakes to share and a higher level of fairness is beneficial to everybody. In particular, b indicates the expected benefit for C due to the aggregate fairness, while e indicates his/her expected private earnings. Given that all these figures are *expected*, the intermediaries' benefits are weighted by the probability p to reach a given level of fairness f. The more C plays fair the more he/she contributes to providing a context for fairness ($p'_f > 0$).

It is worth noting that a crucial component of the model is the intrinsic motivation of subjects (d). This is expected to increase with the overall level of fairness ($d'_f > 0$) and to decline with private material earnings due to crowding-out effects ($d'_e < 0$). Finally, we assumed that a fixed small but strictly positive cost k of fair behaviour existed due the cognitive requirements by C (e.g., information search, memory, and time) to perform a thoughtful evaluation of B.

Assuming that intermediaries choose the level of fairness f^* maximising their expected benefit, we derived the first order condition as follows:

$$p'_f(b+e)+d'_f=0. \qquad (3)$$

It is important to note that intermediaries could see a benefit b from higher levels of fairness only if they are expected to play as A or B in the future, that is, if roles are rotating. If roles are fixed, $b = 0$ and the intermediary's decision depends only on the personal earnings e and the intrinsic motivation d. Given (3), *ceteris paribus* a situation where $b = 0$ is expected to lead to a lower f^* than where intermediaries can derive benefits from the aggregate level of fairness by playing other roles in the future.

It is worth noting that this can be framed in terms of indirect reciprocity [23, 26, 27]. That is the idea of benefiting unknown trustors by punishing self-interested trustees to keep the fairness standards high and benefit from the reciprocity of other reliable intermediaries when cast as trustors (i.e., $b > 0$). This can induce intermediaries to provide reliable information and investors to trust reputational information. This concatenation of strategies cannot work in a system where interaction roles are fixed, given that intermediaries cannot expect benefits from their evaluation as future trustors (i.e., $b = 0$). In this case, trustors cannot expect that intermediaries provide a careful evaluation.

This led us to formulate our first hypothesis as follows.

Hypothesis 1. If roles of the interaction are fixed, indirect reciprocity motives cannot motivate positive behaviour by the parties involved, who will not see future benefits in keeping the levels of fairness high. This implies that ceteris paribus the fixed role condition will decrease fairness and cooperation.

The situation is different when the interests of intermediaries and trustors are aligned. In this case, intermediaries receive a direct payoff $e > 0$ by behaving fair. Indeed, this interest alignment transforms the trust relationship in a typical principal-agent model, where the intermediary (the agent) can behave in the interest of the trustor (the principal). In this case, the rational choice theory predicts that monetary incentives are crucial to motivate the intermediary to act on the trustor's behalf, by guaranteeing that the self-interest of the former coincides with the objectives of the latter [28]. The fact that the negative effect of monetary incentives on intrinsic motivations ($d'_e < 0$) could be compensated by higher levels of fairness ($d'_f > 0$) will lead to higher levels of fairness and cooperation compared to the $e = 0$ case.

This led us to formulate our second hypothesis as follows.

Hypothesis 2. If the intermediary responds to monetary incentives that are aligned with the trustor's interests, higher levels of fairness will lead to higher cooperation in the system.

In the symmetric case, where monetary incentives are aligned with the trustee's interests, intermediaries will receive a personal payoff from being unfair. Note also that, in this case, not only do the incentives crowd out intrinsic motivations but also d declines due to expected lower levels of fairness. This is why the intermediary is in a potential conflict of interest as he/she may be tempted to cheat the trustor by providing opinions that benefit the trustee. Knowing this, the trustor could be induced to question the reliability of the intermediary's opinion and decide not to enter into the exchange or reduce his/her investment to minimal levels. This is expected to erode the basis of cooperation.

This led us to formulate our third hypothesis as follows.

Hypothesis 3. When the intermediary's incentives are aligned with the trustee's interests the levels of fairness will decline, leading to lower cooperation in the system in comparison with the situation where no monetary incentives exist.

A particular case is when intermediaries receive monetary incentives that are independent of their actions and the level of fairness in the system, as in the case of fixed rewards. In this case, e will not enter the benefit calculus as expressed by the first term of (2) as it will be earned in all cases, while the incentive will decrease the intermediaries' intrinsic motivation because of $d'_e < 0$, leading to low levels of fairness. This will make the intermediary's opinion poorly credible for both the trustor and the trustee.

This led us to formulate our fourth hypothesis as follows.

Hypothesis 4. With intermediary's incentives that are fixed and independent from the interests of both the trustor and the trustee, the levels of fairness and cooperation in the system will be lower than in case of no monetary incentives.

Without any material interest in the exchange ($e = 0$), intermediaries intrinsic motivations are expected to increase with f. As long as roles change over time, being $b > 0$ and $p'_f > 0$, subjects are expected to personally benefit from higher levels of fairness in the system. Furthermore, the absence of monetary incentives can transform the interaction into a moral problem, with intermediaries induced to punish misbehaviour by trustees even more than in other incentive schemes [29, 30].

This led us to formulate our last hypothesis as follows.

Hypothesis 5. With alternating roles and without monetary incentives for intermediaries, fairness will increase leading to high levels of cooperation in the system.

3. Methods

To test our hypotheses, we built two experiments based on a modified version of the repeated investment game described above with $d_A = d_B = 10$ MU and $m = 3$ and with the restriction of choices to integer amounts (see the Appendix for a detailed description of the experiments). We extended the original dyadic game towards a third-party game where we introduced intermediaries (Players C) not directly involved in the transaction but asked to rate trustees' behaviour (Players B) for the benefit of the trustors (Players A). The opinion of Players C was formulated individually and was shared with both Players A and B involved in the exchange. When selected as a C Player, the subject was matched with one Player A and one Player B and privately informed of the amount received and returned by the latter in the previous period. Then, he/she was asked to rate Player B's behaviour as "negative," "neutral," or "positive." His/her opinion was privately displayed to Player A before his/her investment decision. The rest of the game worked as the standard dyadic version described above. Note that any communication between subjects was forbidden and subjects played anonymously with possible partners from a large pool of subjects.

In the first experiment, game roles (i.e., trustor, trustee, and intermediary) alternated regularly throughout the game, with the same subject playing the same number of times in the three roles with randomly determined partners [11]. The second experiment followed the same design except that roles were fixed throughout the game. The aim of this second experiment was to rule out any (indirect) reciprocity motive from the intermediary behaviour, as the intermediary now could not provide reliable information to expect good future information in turn (in the trustor's role).

More specifically, in both experiments, monetary incentives for intermediaries systematically varied across treatments according to our hypotheses. In the *No incentive* treatment, intermediaries did not receive any rewards for their task, also losing potential earnings as trustors or trustees when selected to play in this role. In the *Fixed incentive* treatment, intermediaries received a fixed payoff of 10 MU, equal to the trustor and trustee endowments. In the A *incentive* treatment, intermediaries earnings were equal to the payoff obtained by the trustors they advised. In the B *incentive* treatment, intermediaries' earnings were equal to the payoff obtained by the trustees they rated.

4. Results

4.1. Experiment 1 (Alternating Roles). Our experiment produced investments and returns comparable to previous experiments in the *Baseline* [18] and, consistent with previous studies which introduced reputational motives [9, 22], showed that the presence of the intermediaries dramatically improved cooperation.

A total of 136 subjects (50% female) participated in the experiment, held in late 2010. Both investments and returns were higher when intermediaries were introduced, with investments increasing from an average of 3.22 MU in the *Baseline* up to 5.21 MU in A *incentive* and returns rising from 2.00 in the *Baseline* to 6.87 in *No incentive* (Figure 1). (Our dataset may be accessed upon request to the corresponding author. All statistical analyses were performed using the *R* 2.15.1 platform [31]. Please, note that the amounts exchanged in the first three periods of the game, when intermediaries had no previous information to evaluate, and in the last three periods, when trustees knew that no further rating would take place, were not included in the analysis.)

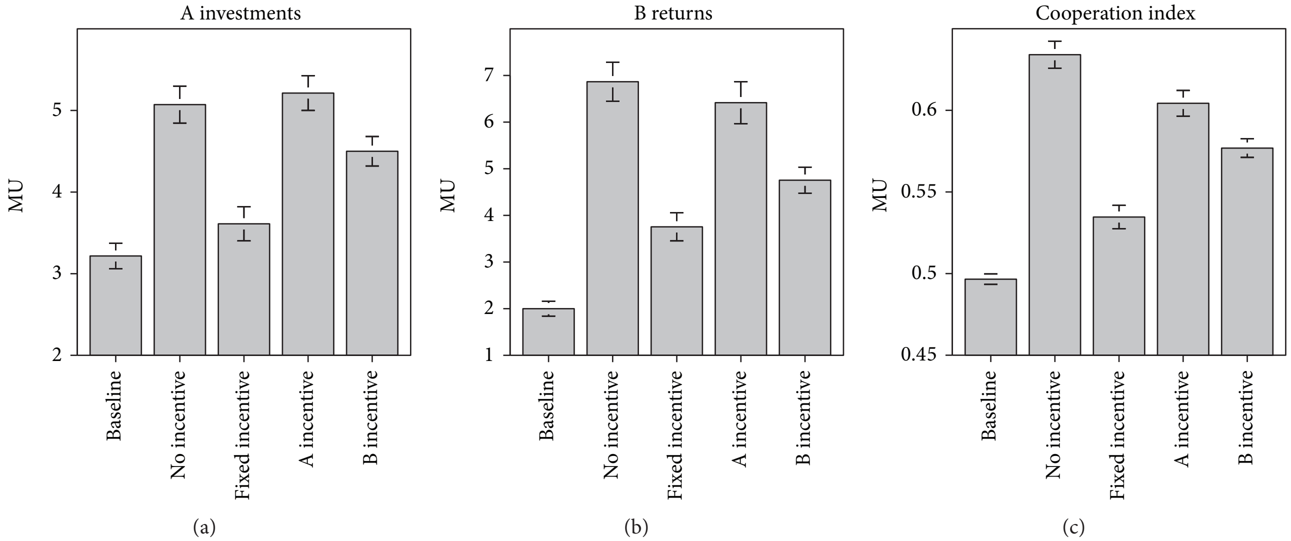

Figure 1: Experiment 2 results. Average investments, returns, and CI by treatment. Bars show the standard error of the means.

Differences with the *Baseline* for both investments and returns were significant at the 5% level for all treatments except *Fixed incentive*, where the difference was significant only for returns. Significant differences also existed for Player B returns. Both *No incentive* and A *incentive* led to higher returns than *Fixed incentive* (Wilcoxon rank sum tests on individual averages, $W = 531.0$, $P = 0.002$ one tailed, and $W = 199.0$, $P = 0.002$ one tailed, resp.). There were no significant differences between *No incentive* and A *incentive* ($W = 385.0$, $P = 0.365$). Differences in investments were not considerable but still statistically significant at 5% between *No incentive* and *Fixed incentive* ($W = 508.0$, $P = 0.006$ one tailed) and between A *incentive* and *Fixed incentive* ($W = 176.5$, $P = 0.001$ one tailed).

To compare the outcome of the different treatments better, we built an indicator that considered both social welfare and fairness, which are two aspects strictly linked with cooperation due to the social dilemma nature of the investment game [32]. Indeed, in any investment game, trust is a double-edged sword. On the one hand, it includes trustors who are better-off by placing trust and so can increase the welfare of the system by taking the risk of interaction. On the other hand, it also includes trustees who can act in a more or less trustworthy way, so contributing to giving rise to a fair exchange. Following [11], in order to look at these two sides of the this process, we built an index that combined attention to the social welfare, which only depended on A investments, and fairness, which depended on B returns, which we called *cooperation index* (CI). The level of social welfare is an important indicator of the system efficiency in the different treatments. This is indicated by $E = I/d_A$, where I was A investment and d_A was the endowment. This indicator took 0 when A invested nothing and 1 when A invested the whole endowment. Following previous research [33–36], *fairness* was calculated by comparing the difference in the payoffs to the sum of monetary gains: $F = 1 - [|P_A - P_B|/(P_A + P_B)]$, where P_A and P_B were the payoffs earned by A and B Players, respectively. This was 0 when one of the players obtained the whole amount at stake and the other received nothing, while it was 1 when both players obtained the same payoff. Note that no trade-off exists between maximising E and F, given that, for any level of investment (including zero), only the amount returned determined the fairness payoff. The cooperation index was defined as $\mathrm{CI} = (E + F)/2$. This was 0 when A Players invested nothing and B Players returned all their endowments, grew with the growth of A investments and a fairer distribution of final payoffs, and became 1 when A Players invested d_A and B Players returned half of their total endowment, that is, $(d_B + mI)/2$.

The treatment with the highest CI was *No incentive*, which led to more fairness than any other treatment (Figure 1). Differences in the CI were statistically significant at 10% with A *incentive* and at 5% with all other treatments. The high CI value in *No incentive* was especially important as in this case, unlike A *incentive*, intermediaries had no monetary incentive to cooperate with trustors. This would confirm that the lack of monetary incentives for intermediaries implied higher normative standards of behaviour for the other actors involved.

4.2. Experiment 2 (Fixed Roles). A total of 244 subjects (55% females) participated in the second experiment, which was organized in 2011. Results showed that trust and trustworthiness were generally lower than in the first experiment. Average investments ranged from 2.36 MU in *Fixed incentive* to 3.96 MU in A *incentive*. Returns ranged from 2.47 MU in *Fixed incentive* to 5.27 MU in A *incentive*. The cooperation index ranged from 0.507 MU in the *Baseline* to 0.557 MU in A *incentive* (Figure 2).

The only treatment leading to investments significantly higher than the *Baseline* was A *incentive* ($W = 82$, $P = 0.049$ one tailed), while returns were significantly higher (at the 10% level) in both A *incentive* and B *incentive* ($W = 88$, $P = 0.071$

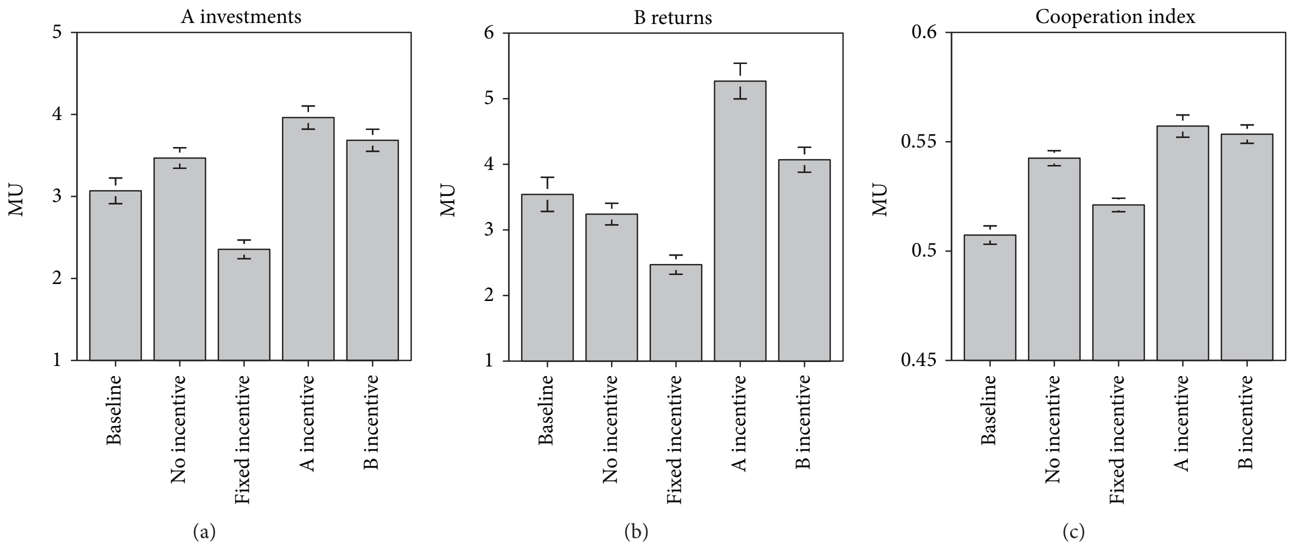

Figure 2: Experiment 2 results. Average investments, returns, and CI by treatment. Bars show the standard error of the means.

one tailed for both treatments). Finally, the cooperation index was significantly higher than the *Baseline* in *No incentive* ($W = 351$, $P = 0.019$ one tailed), A *incentive* ($W = 299$, $P = 0.003$ one tailed) and B *incentive* ($W = 312$, $P = 0.004$), but not in *Fixed incentive* ($W = 462$, $P = 0.288$ one tailed).

4.3. Comparing the Two Experiments. A comparison between alternating versus fixed role treatments highlighted the effects of indirect reciprocity (Table 1). Besides lower returns in the *Baseline*—a result consistent with the existing literature [18]—investments and returns were generally higher in the alternating roles experiment. This difference was especially relevant in *No incentive*. This was one of the best performing treatments in the first experiment, while, in fixed role treatments, investments and returns were similar to the *Baseline* (i.e., without intermediaries). On the other hand, the two treatments with sound monetary incentives led to a more modest decrease in investment and especially, in returns. Note also that B *incentive* performed similarly in the two experiments.

In order to examine the effect of the different factors involved in the exchange, we estimated a random effects model using dummies indicating the fixed role experiment, each treatment, the first and the last period, and the second half of the game as regressors (Table 2). Due to the considerable interdependence between the observed decisions, random effects (multilevel) regression analysis was performed that considered the nestedness of observations at the individual level. Results showed that fixed roles caused an overall decrease of trust but had less effect on trustworthiness. A *incentive* had a stronger impact on trustee behaviour, while *No incentive* had a stronger impact on CI, that is, on the fairness of the exchange. All conditions except *Fixed incentives* led to higher trust and trustworthiness than the *Baseline*. Finally, all cooperation indicators declined during the game, confirming the typical “end effect” found in previous studies [9, 22, 37].

If we consider the behaviour of the intermediaries, we can see that they generally played fairly, asking B Players to return significantly more to award a more positive rating (Table 3(a)). They were more demanding where trust was higher, namely, in *No incentive* and A *incentive* of the alternating roles experiment. In both cases, to award a positive rating to trustees, intermediaries asked trustees to return about one-third of the trustees’ endowments. On the other hand, in less cooperative treatments such as *No incentive* or *Fixed incentive* in the fixed roles experiment, they were less demanding (i.e., to one-sixth and even less). Moreover, negative ratings were more frequent in the most cooperative treatments. Although only descriptive, these results suggest that a more rigorous reputation process took place in these conditions, despite the overall higher trustworthiness of trustees achieved in these treatments (Table 3(b)).

It is also important to note that trustor investments reflected the behaviour of the intermediary. Generally, trustors invested more when their opponents received positive ratings, less in cases of neutral ratings and even less in cases of negative ratings (Table 4). It is worth noting that also in this case there were no important differences across treatments. Results suggest that trustors trusted intermediaries more when the latter gave positive ratings in *No incentives* and A *incentive* in the alternating roles experiment. In these cases, they invested on average 6.25 MU and 7.04 MU, respectively. They invested less in treatments where there was low trust, such as *Fixed incentive* in the fixed role condition, where average investment of trustors was 2.43 MU even when trustees had a positive rating.

5. Discussion

Our experiments have highlighted the crucial role of independent judges in trust situations, even when their opinions could be viewed as subjective and/or without serious long-term future consequences for the trustee. We found that,

Table 1: Overview of Experiments 1 and 2 results. The AR/FR column presents the ratio between the corresponding results in the alternating and the fixed role experiment. The last two columns present Wilcoxon rank sum tests of the null hypothesis that the result distribution is the same in the two experiments.

Treatment		Altern. roles		Fixed roles		AR/FR	Wilcoxon	
		Mean	SE	Mean	SE		W	P
Baseline	A investments	3.22	0.16	3.07	0.16	1.05	207.5	0.385
	B returns	2.00	0.16	3.54	0.26	0.56	142.5	0.079
	B ret. (prop.)	0.09	0.01	0.16	0.01	0.56	137.0	0.059
	CI	0.50	0.00	0.51	0.00	0.98	319.0	0.118
No incentive	A investments	5.07	0.23	3.47	0.12	1.46	174.0	0.029
	B returns	6.87	0.42	3.24	0.16	2.12	196.0	0.003
	B ret. (prop.)	0.24	0.01	0.14	0.01	1.71	189.0	0.006
	CI	0.63	0.01	0.54	0.00	1.17	396.0	0.000
Fixed inc.	A investments	3.61	0.21	2.36	0.11	1.53	170.0	0.040
	B returns	3.75	0.30	2.47	0.15	1.52	142.0	0.232
	B ret. (prop.)	0.17	0.01	0.10	0.01	1.70	150.0	0.153
	CI	0.54	0.01	0.52	0.00	1.04	264.0	0.319
A incentive	A investments	5.21	0.21	3.96	0.14	1.32	174.5	0.028
	B returns	6.42	0.45	5.27	0.27	1.22	147.0	0.180
	B ret. (prop.)	0.23	0.01	0.21	0.01	1.10	134.0	0.333
	CI	0.60	0.01	0.56	0.01	1.07	365.0	0.002
B incentive	A investments	4.50	0.18	3.68	0.14	1.22	154.0	0.121
	B returns	4.75	0.28	4.07	0.19	1.17	141.5	0.238
	B ret. (prop.)	0.19	0.01	0.18	0.01	1.06	130.0	0.387
	CI	0.58	0.01	0.55	0.00	1.05	307.0	0.071

Table 2: Multilevel regression analysis with individual-level random effects. Significance codes: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$, $^{\dagger}P < 0.1$.

Dependent	Investments	Returns	CI
(Intercept)	3.607^{***}	0.154	0.516^{***}
Fixed roles	-0.962^{***}	-0.133	-0.030^{***}
No incentive	1.339^{**}	1.509^{**}	0.090^{***}
Fixed incentive	0.040	0.684	0.030^{*}
A incentive	1.538^{***}	2.004^{***}	0.083^{***}
B incentive	1.075^{**}	$1.012^{\dagger}$	0.066^{***}
First period	0.066	0.954^{***}	0.024^{*}
Last period	-0.562^{**}	-0.848^{**}	-0.031^{**}
Periods 16–30	-0.313^{***}	-0.420^{***}	-0.010^{*}
A investment		0.853^{***}	
Random effects			
σ (id)	1.827	2.553	0.054
σ (residual)	2.314	2.971	0.118
Number of interactions	4070	4070	4070
Number of individuals	222	222	222
F	21.5^{***}	320.7^{***}	121.1^{***}

somewhat counterintuitively, the intermediary's opinions had a stronger effect when the intermediary had no material interest from the exchange. If there was room for indirect reciprocity between the parties involved, triadic relationships could provide a moral base to overcome the negative traps of self-interest. We also found that monetary incentives and moral action were not clearly separable or substitutes [38]. If properly designed, monetary incentives for intermediaries had a positive effect, especially on trustor's investments, but were less effective than social norms and reciprocity in ensuring fair exchanges in terms of more equal distribution of payoffs between trustors and trustees.

First, the "shadow of reciprocity" implied that intermediaries kept the standards of evaluation high to benefit good information in turn by other intermediaries, when acting as trustors. This in turn induced trustees to be more reliable. Under evaluation, trustees overestimated the future impact of their good standing and behaved more fairly. On the other hand, having information about trustee behaviour induced trustors to predict higher cooperative responses by trustees and increase their investment. It is worth noting that recent work showed that this mechanism is relatively independent of the quality of information. For instance, in an experiment on financial decisions in situations of uncertainty and information asymmetry similar to our experiment, [39] showed that the availability of information from other subjects increased risky trust investment decisions of subjects independent of the quality of the information.

It is worth noting that the motivation crowding-out model (2) was able to match our empirical data qualitatively. Coherently with H1, the comparison of the two experiments confirmed the paramount importance of indirect reciprocity motives in trust situations. The idea that indirect reciprocity is fundamental in human societies has been suggested by [40],

Table 3: Summary statistics by rating. Average return proportion by rating (a) and rating distribution per treatment (b). AR stands for "alternating roles" and FR for "fixed roles."

(a) Return proportion

Treatment	Negative	Neutral	Positive
No incentive AR	0.14	0.22	0.34
No incentive FR	0.13	0.15	0.14
Fixed incentive AR	0.08	0.16	0.24
Fixed incentive FR	0.11	0.09	0.17
A incentive AR	0.15	0.22	0.33
A incentive FR	0.18	0.16	0.28
B incentive AR	0.14	0.20	0.22
B incentive FR	0.19	0.15	0.18
All treatments	0.15	0.16	0.22

(b) Rating distribution

Treatment	Negative	Neutral	Positive
No incentive AR	0.42	0.22	0.36
No incentive FR	0.37	0.26	0.37
Fixed incentive AR	0.34	0.31	0.35
Fixed incentive FR	0.36	0.30	0.34
A incentive AR	0.41	0.25	0.35
A incentive FR	0.37	0.23	0.40
B incentive AR	0.33	0.29	0.38
B incentive FR	0.33	0.22	0.44
All treatments	0.36	0.26	0.38

Table 4: Average investments by treatment and rating.

Treatment	Negative	Neutral	Positive
No incentive AR	3.72	5.76	6.25
No incentive FR	2.39	3.24	4.62
Fixed incentive AR	1.84	4.05	4.81
Fixed incentive FR	2.18	2.44	2.43
A incentive AR	3.43	5.02	7.04
A incentive FR	2.75	4.46	4.71
B incentive AR	3.39	4.52	5.27
B incentive FR	2.79	4.00	4.10
All treatments	2.72	3.87	4.56

who argued that this is one of the most crucial forces in human evolution. This has been found in different experimental settings [21, 23]. Our experiments suggest that, *ceteris paribus*, indirect reciprocity explains a significant part of cooperation also in triadic relations (Tables 1 and 2). This indicates that evaluation systems where the roles of trustors, trustees, and evaluators rotate could work better and more efficiently than those in which roles are fixed.

H2 was also confirmed by our data. In both experiments A *incentive* guaranteed high levels of cooperation as intermediary evaluations fostered both investments by trustors and returns by trustees. In this treatment, trustees followed reputation building strategies and so returned more. Furthermore, trustors considered the opinions of intermediaries credible and used them to discriminate between "good" and "bad" opponents. Therefore, intermediaries were functional to encapsulate mutual trust and cooperation.

H3 was the only hypothesis not fully supported by our data. On the one hand, B *incentive* led to less cooperation than A *incentive*. On the other hand, B *incentive* unexpectedly gave better results than the *Baseline*, as intermediaries were less demanding than in A *incentive* to award positive ratings but still discriminated between trustworthy and untrustworthy B Players (Table 3). This induced trustors to consider the opinion of reliable intermediaries and followed their ratings independent of the misalignment of mutual interests caused by the monetary incentives (Table 4). This would again confirm that, in condition of information asymmetry, any information on trustee is better than none, as in the *Baseline*.

The mismatch between monetary incentives and intermediary's behaviour is interesting. Results indicated that intermediaries did not predictably respond to incentives as they played fairly in each treatment (Table 3). This can be explained in terms of intrinsic motivations and the sense of responsibility that are typically associated with such "neutral" positions. More specifically, it is worth noting that willingness to provide pertinent judgement by intermediaries was not sensitive to any variations in the incentive scheme. This would testify to the inherent moral dimension of this role. On the other hand, even the perception that their task was indirectly judged by the trustors, who were informed of their ratings, could have motivated the intermediaries to take their role seriously, independent of the incentives. More importantly, the alternating roles protocol allowed intermediaries to follow indirect reciprocity strategies and so they kept the credibility and quality of their ratings high in order to benefit from reciprocity by other intermediaries when cast as trustors. This explains the difference between the two experiments.

Consistent with H4, *Fixed incentive* led to less cooperation than other schemes. A comparison between *No incentive* and *Fixed incentive* allows us to understand this point better. In the alternating roles experiment, while *No incentive* led to more trust and cooperation, *Fixed incentive* barely improved the *Baseline*, that is, when intermediaries were not present. This does not make sense in a rational choice perspective, as in both cases the incentives of intermediaries were ambiguous. On the other hand, while intrinsic motivations were crucial to induce intermediaries to formulate reliable ratings in *No incentive*, these aspects were crowded out by monetary incentives in *Fixed incentive*. This is consistent with the *motivation crowding theory* [25, 41] and with many empirical studies that showed that incentivization policies which targeted self-interested individuals actually backfire by undermining individual "moral sentiments" in a variety of social and economic situations [29, 38]. In this sense, some possible extensions of our work could examine the relationship between crowding out effects on social norms and the presence of different incentive schemes in more detail. For instance, it would be interesting to understand if different incentives for the intermediaries could influence future interactions between trustors and trustees without intermediaries. This could also help us to look at self-reinforcing effects of incentives and social norms on trust.

It is worth considering that more consistent monetary incentives for intermediaries could increase their credibility for the other parties involved, so motivating more cooperation. Indeed, a possible explanation of the low cooperation in *Fixed incentive* is that the magnitude of our monetary incentives was sufficient to crowd out intrinsic motives of subjects without promoting reciprocal and self-interested behaviour, as what happened in typical monetary markets [29]. Further work is necessary to examine this hypothesis by testing fixed incentives of different magnitudes, although it is worth considering that there are constraints in terms of magnitude of incentives that can be implemented also in real situations.

H5 was also confirmed by our results, at least in the alternating roles experiment. In this case, *No incentive* was the best treatment for cooperation. On the other hand, it did not promote trust and especially, trustworthiness when roles were fixed. As argued above, by fixing the roles, there was no room for indirect reciprocity strategies. The fact that intermediaries could expect future benefits from their roles and were subsequently cast as both trustors and trustees induced the parties involved to believe more in the credibility of the intermediaries' opinion. Also in this case, further empirical investigation is needed to compare social situations where intermediaries have a specialised role with situations where there is voluntarism and mixture of roles.

These findings imply that the "mantra" of incentivization popularized by most economists as a means to solve trust and cooperation problems, especially in economic and public policy, should be seriously reconsidered [42]. Not only should incentives be properly designed to produce predictable outcomes and this is often difficult, but also incentive-response behaviour of individuals is more heterogeneous and unpredictable than expected [43]. If this occurred in a simple laboratory game, where individuals had perfect information and the rules of the game were fully intelligible, one should expect even more heterogeneity of individual behaviour and unpredictability of social outcomes in real situations.

Our results finally suggest that insisting on incentivization potentially crowds out other social norms-friendly, endogenous mechanisms such as reputation, which could ensure socially and economically consistent results. While incentives might induce higher investment risks (but only if properly designed), social norms can also help to achieve a fairer distribution, nurturing good behaviour which can be even more endogenously sustainable in the long run [44].

Appendices

A. Details of the Experiment Organization

All participants were recruited using the online system ORSEE [45]. They were fully informed and gave their consent when they voluntarily registered on ORSEE. Data collection fully complied with Italian law on personal data protection (D.L. 30/6/2003, n. 196). Under the applicable legal principles on healthy volunteers' registries, the study did not require ethical committee approval. All interactions were anonymous and took place through a computer network equipped with the experimental software z-Tree [46].

Experiment 1 (Alternating Roles). A total of 136 subjects (50% female) participated in the experiment held at the GECS experimental lab (see http://www.eco.unibs.it/gecs/) in the late 2010. Each experimental session took less than one hour and participants earned, on average, 14.82 Euro, including a 5-Euro show-up fee.

Twenty-eight subjects participated in a *Baseline* repeated investment game (hereafter *Baseline*) set using the following parameters: $d_A = d_B = 10$ monetary units (MU), $m = 3$. Each MU was worth 2.5 Euro Cents and subjects were paid in cash immediately at the end of the experiment. The game was repeated 30 times with couples who were randomly reshuffled in each period. Players' roles regularly alternated throughout the game. This meant that each subject played exactly 15 times as A and 15 times as B.

All the other treatments, each played by 27 subjects, introduced a third player into the game (Player C) in the role of intermediary. Once C Players were introduced, we varied the monetary incentive schemes offered to them. In the *No incentive* treatment, intermediaries did not receive any rewards for their task, also losing potential earnings as trustors or trustees when selected to play in this role. In the *Fixed incentive* treatment, intermediaries received a fixed payoff of 10 MU, equal to the trustor and trustee endowments. In the A *incentive* treatment, intermediaries earnings were equal to the payoff obtained by the trustors they advised. In the B *incentive* treatment, intermediaries' earnings were equal to the payoff obtained by the trustees they rated.

Experiment 2 (Fixed Roles). A total of 244 subjects (55% female) participated in the second experiment, which was organized at the GECS experimental lab in the late 2011. Participants were recruited and played as in the first experiment. Each experimental session took less than one hour and participants earned, on average, 14.78 Euro, including a 5-Euro show-up fee.

Any overlap with the first experiment participants was avoided, so that the two experiments could in principle be viewed as a single experiment with a between-subject design. The treatments were as before, with the only difference that roles remained fixed throughout the game. Note that the fact that roles no longer alternated actually reduced the sample of observations per role, with a considerable consequence especially on the treatments involving three parties (i.e., all but the *Baseline*). To overcome this problem, we doubled the number of subjects participating in *No incentive*, *Fixed incentive*, A *incentive*, and B *incentive*, which were played by 54 subjects each, organized in two sessions, each one involving 27 participants.

B. Instructions

Before the beginning of the game, participants read the game instruction on their computer screen. Subsequently, they filled a short test designed to check their understanding of the game. Participants could also ask the experimenters any further questions or issues. Here is the English translation of the original Italian instructions. Sentences in italics are treatment or experiment specific, whereas a normal font is used for instructions common to all treatments/experiments.

Screen 1: Overall Information on the Experiment

(i) All these instructions contain true information and are the same for all participants.

(ii) Please, read them very carefully. At the end, some questions will be asked by the system to test your understanding of the experiment.

(iii) The experiment concerns economic problems.

(iv) During the experiment, you will be asked to take decisions, upon which your final earnings will depend. Earnings will be paid in cash at the end of the experiment.

(v) Each decision will take place through your computer screen.

(vi) During the experiment, it is prohibited to talk with anyone. If you do so, you will be excluded from the experiment and you will lose your earnings. Please, turn your mobile phones off.

(vii) For any information and question, put your hands up and wait until an experimenter comes to your position.

(viii) During the experiment, virtual monetary units (MU) are used that have a fixed exchange rate with real Euros. For each MU earned in the experiment, you will receive 2.5 Euro Cents. For example, if at the end of the experiment, your earning is 500 MU, this means that you will receive 12.50 Euros, plus a fixed show-up fee of 5 Euros.

Screen 2: Interaction Rules

(i) The experiment consists of a sequence of interaction rounds between pairs of players (*Baseline*).

(ii) The experiment consists of a sequence of interaction rounds between groups of three players (*all three person treatments*).

(iii) Pairs are randomly matched and change each round; therefore, they are made up of different individuals each round (*Baseline*).

(iv) Groups are randomly matched and change each round; therefore, they are made up of different individuals each round (*all three person treatments*).

(v) There is no way to know who you are playing with, nor is it possible to communicate with her/him.

(vi) In each pair, each participant will perform a different role; roles are called "Player A" and "Player B" (*Baseline*).

(vii) In each group, each participant will perform a different role; roles are called "Player A," "Player B," and "Player C" (*all three person treatments*).

(viii) Roles are randomly assigned at the beginning of the experiment and then regularly changed for the rest of it. For example, one participant in the pair will play a sequence of rounds such as Player A, Player B, Player A, Player B, Player A, and Player B, and the other one will play a sequence such as Player B, Player A, Player B, Player A, Player B, and Player A. Therefore, over the experiment, all participants will play both roles the same number of rounds (*alternating roles experiment, Baseline*).

(ix) Roles are randomly assigned at the beginning of the experiment and then regularly changed for the rest of it. For example, one participant in the group will play a sequence of rounds such as Player A, Player B, Player C, Player A, Player B, and Player C, and another one will play a sequence such as Player B, Player C, Player A, Player B, Player C, and Player A. Therefore, over the experiment, all participants will play all roles the same number of rounds (*alternating roles experiment, all three person treatments*).

(x) Roles are randomly assigned at the beginning of the experiment. This means that each participant will play in the same role throughout the whole experiment (*fixed roles experiment, all treatments*).

(xi) Each participant should make one decision each round.

(xii) The experiment lasts 30 rounds.

Screen 3: Task Structure

(i) Player A plays first and receives an endowment of 10 MU. She/he has to decide how much of it to keep for her/himself and how much to send to Player B. Player A can send to Player B any amount, from 0 MU to the whole endowment (=10 MU).

(ii) The amount of MU kept by Player A is part of her/his earning. The amount of MU sent to Player B is tripled and assigned to Player B.

(iii) Player B also receives an endowment of 10 MU, as Player A did.

(iv) Example 1: if Player A sends 2 MU, Player B receives 2 × 3 + 10 = 16 MU.

(v) Example 2: if Player A sends 5 MU, Player B receives 5 × 3 + 10 = 25 MU.

(vi) Example 3: if Player A sends 8 MU, Player B receives 8 × 3 + 10 = 34 MU.

(vii) Player B should decide how much of the whole amount received to send to Player A; Player B can send any amount to Player A from 0 MU to the entire sum.

(viii) The amount of MU kept by Player B represents her/his earning; the amount of MU sent to Player A will add up to Player A earning.

(ix) Before A and B players, respectively, take their decisions, Player C is asked to rate the decision Player B took the previous round (since no decision has been taken by B in the first round yet, C is not asked to rate in the first round) (*all three person treatments*).

(x) Player C can rate Player B's decision as "negative," "neutral," or "positive" and the rating is communicated to both A and B Players with whom he/she is grouped (in the first rounds, since there is no decision undertaken by Player B yet, the rating of Player B is assigned by the system as "unknown") (*all three person treatments*).

(xi) Player C does not receive any earnings for her/his action (*No incentive treatments*).

(xii) Player C earning is fixed and is equal to 10 MU (*Fixed incentive treatments*).

(xiii) Player C earning is equal to what Player A will earn at the end of the round (A *incentive treatments*).

(xiv) Player C earning is equal to what Player B will earn at the end of the round (B *incentive treatments*).

(xv) Earnings will be added up each round to give the final reward for each participant, which will be paid at the end of the experiment.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to gratefully acknowledge comments on a preliminary version of this paper by Karl-Dieter Opp, Gianluca Carnabuci, Michael Maes, and the audience at the INAS Conference on Analytical Sociology, Columbia University, New York, 8-9 June, 2012, and at the ASA 2012 Annual Meeting, Denver, 17–20 August, 2012. They also would like to thank Marco Castellani and Niccolò Casnici for help during the experiment and Fabio Recinella for technical assistance in the lab. This work was supported by the Collegio Carlo Alberto and the University of Brescia. The authors also gratefully acknowledge support from the Intra-European Fellowship Program of the European Union (Grant no. GA-2009-236953).

References

[1] J. S. Coleman, *Foundations of Social Theory*, Harvard University Press, Harvard, Mass, USA, 1990.

[2] R. S. Burt and M. Knez, "Kinds of third-party effects on trust," *Rationality and Society*, vol. 7, no. 3, pp. 255–292, 1995.

[3] K. S. Cookm and R. Hardin, "Norms of cooperativeness and networks of trust," in *Social Norms*, M. Hechter and K.-D. Opp, Eds., pp. 327–347, Russel Sage Foundation, New York, NY, USA, 2001.

[4] M. Bateson, D. Nettle, and G. Roberts, "Cues of being watched enhance cooperation in a real-world setting," *Biology Letters*, vol. 2, no. 3, pp. 412–414, 2006.

[5] I. Bohnet and B. S. Frey, "The sound of silence in prisoner's dilemma and dictator games," *Journal of Economic Behavior and Organization*, vol. 38, no. 1, pp. 43–57, 1999.

[6] T. N. Cason and V.-L. Mui, "A laboratory study of group polarisation in the team dictator game," *Economic Journal*, vol. 107, no. 444, pp. 1465–1483, 1997.

[7] K. J. Haley and D. M. T. Fessler, "Nobody's watching? Subtle cues affect generosity an anonymous economic game," *Evolution and Human Behavior*, vol. 26, no. 3, pp. 245–256, 2005.

[8] M. Rigdon, K. Ishii, M. Watabe, and S. Kitayama, "Minimal social cues in the dictator game," *Journal of Economic Psychology*, vol. 30, no. 3, pp. 358–367, 2009.

[9] R. Boero, G. Bravo, M. Castellani, and F. Squazzoni, "Reputational cues in repeated trust games," *Journal of Socio-Economics*, vol. 38, no. 6, pp. 871–877, 2009.

[10] M. Bacharach and D. Gambetta, "Trust in signs," in *Trust and Society*, K. S. Cook, Ed., vol. 2, pp. 148–184, Russell Sage, New York, NY, USA, 2001.

[11] F. Squazzoni, G. Bravo, and K. Takács, "Does incentive provision increase the quality of peer review? An experimental study," *Research Policy*, vol. 42, no. 1, pp. 287–294, 2013.

[12] K. Cook, R. Hardin, and M. Levi, *Cooperation without Trust*, Russell Sage, New York, NY, USA, 2005.

[13] R. Hardin, *Trust and Trustworthiness*, Russell Sage, New York, NY, USA, 2004.

[14] J. P. Bailey and Y. Bakos, "An exploratory study of the emerging role of electronic intermediaries," *International Journal of Electronic Commerce*, vol. 1, no. 3, pp. 7–20, 1997.

[15] J. W. Palmer, J. P. Bailey, and S. Faraj, "The role of intermediaries in the development of trust on the WWW: the use and prominence of trusted third parties and privacy statements," *Journal of Computer-Mediated Communication*, vol. 5, no. 3, 2000.

[16] O. Sorenson and T. E. Stuart, "Syndication networks and the spatial distribution of venture capital investments," *American Journal of Sociology*, vol. 106, no. 6, pp. 1546–1588, 2001.

[17] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity and social history," *Games and Economic Behavior*, vol. 10, no. 1, pp. 122–142, 1995.

[18] N. D. Johnson and A. A. Mislin, "Trust games: a meta-analysis," *Journal of Economic Psychology*, vol. 32, no. 5, pp. 865–889, 2011.

[19] V. Buskens and W. Raub, "Rational choice research on social dilemmas: embeddedness effects on trust," in *Handbook of Rational Choice Social Research*, R. Wittek, T. A. B. Snijders, and V. Nee, Eds., Russell Sage, New York, NY, USA, 2008.

[20] V. Buskens, W. Raub, and J. van der Veer, "Trust in triads: an experimental study," *Social Networks*, vol. 32, no. 4, pp. 301–312, 2010.

[21] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, no. 6960, pp. 785–791, 2003.

[22] C. Keser, "Experimental games for the design of reputation management systems," *IBM Systems Journal*, vol. 42, no. 3, pp. 498–506, 2003.

[23] I. Seinen and A. Schram, "Social status and group norms: indirect reciprocity in a repeated helping experiment," *European Economic Review*, vol. 50, no. 3, pp. 581–602, 2006.

[24] D. Barrera and G. G. Van De Bunt, “Learning to trust: networks effects through time,” *European Sociological Review*, vol. 25, no. 6, pp. 709–721, 2009.

[25] B. S. Frey and F. Oberholzer-Gee, “The cost of price incentives: an empirical analysis of motivation crowding- out,” *American Economic Review*, vol. 87, no. 4, pp. 746–755, 1997.

[26] R. D. Alexander, *The Biology of Moral Systems*, Basic Books, New York, NY, USA, 1987.

[27] M. A. Nowak and K. Sigmund, “The dynamics of indirect reciprocity,” *Journal of Theoretical Biology*, vol. 194, no. 4, pp. 561–574, 1998.

[28] J.-J. Laffont and D. Martimort, *The Theory of Incentives: The Principal-Agent Model*, Princeton University Press, Princeton, NJ, USA, 2002.

[29] D. Ariely, A. Bracha, and S. Meier, “Doing good or doing well? Image motivation and monetary incentives in behaving prosocially,” *American Economic Review*, vol. 99, no. 1, pp. 544–555, 2009.

[30] J. Heyman and D. Ariely, “Effort for payment: a tale of two markets,” *Psychological Science*, vol. 15, no. 11, pp. 787–793, 2004.

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012.

[32] P. Kollock, “Social dilemmas: the anatomy of cooperation,” *Annual Review of Sociology*, vol. 24, pp. 182–214, 1998.

[33] I. Almås, A. W. Cappelen, E. Ø. Sørensen, and B. Tungodden, “Fairness and the development of inequality acceptance,” *Science*, vol. 328, no. 5982, pp. 1176–1178, 2010.

[34] E. Fehr and K. M. Schmidt, “A theory of fairness, competition, and cooperation,” *Quarterly Journal of Economics*, vol. 114, no. 3, pp. 817–868, 1999.

[35] M. A. Nowak, K. M. Page, and K. Sigmund, “Fairness versus reason in the ultimatum game,” *Science*, vol. 289, no. 5485, pp. 1773–1775, 2000.

[36] M. Rabin, “Incorporating fairness into game theory and economics,” *American Economic Review*, vol. 83, pp. 1281–1302, 1993.

[37] F. Cochard, P. Nguyen Van, and M. Willinger, “Trusting behavior in a repeated investment game,” *Journal of Economic Behavior and Organization*, vol. 55, no. 1, pp. 31–44, 2004.

[38] S. Bowles, “Policies designed for self-interested citizens may undermine “the moral sentiments”: evidence from economic experiments,” *Science*, vol. 320, no. 5883, pp. 1605–1609, 2008.

[39] R. Boero, G. Bravo, M. Castellani, and F. Squazzoni, “Why bother with what others tell you? An experimental data-driven agent-based model,” *Journal of Artificial Societies and Social Simulation*, vol. 13, no. 3, article 6, 2010.

[40] M. A. Nowak and R. Highfield, *SuperCooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*, Free Press, New York, NY, USA, 2011.

[41] B. S. Frey and R. Jegen, “Motivation crowding theory,” *Journal of Economic Surveys*, vol. 15, no. 5, pp. 589–611, 2001.

[42] F. Squazzoni, “A social science-inspired complexity policy: beyond the mantra of incentivization,” *Complexity*, vol. 19, pp. 5–13, 2014.

[43] E. Shafir, *The Behavioral Foundations of Public Policy*, Princeton University Press, Princeton, NJ, USA, 2013.

[44] S. Bowles and S. Polanía-Reyes, “Economic incentives and social preferences: substitutes or complements?” *Journal of Economic Literature*, vol. 50, no. 2, pp. 368–425, 2012.

[45] B. Greiner, “An online recruitment system for economic experiments,” in *Forschung und Wissenschaftliches Rechnen 2003*, K. Kremer and V. Macho, Eds., pp. 79–93, Gesellschaft für wissenschaftliche Datenverarbeitung, Göttingen, Germany, 2004.

[46] U. Fischbacher, “Z-Tree: zurich toolbox for ready-made economic experiments,” *Experimental Economics*, vol. 10, no. 2, pp. 171–178, 2007.