Economic Growth Centre
Economics, School of Social Sciences
Nanyang Technological University
14 Nanyang Drive
Singapore 637332

LOYALTY WITHOUT TRUST

Renaud Foucart
Jonathan H.W. Tan

October 2019

EGC Report No: 2019/03

HSS-04-90A
Tel: +65 67906073
Email: D-EGC@ntu.edu.sg

Economic Growth Centre Working Paper Series

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Economic Growth Centre
School of Social Sciences

Loyalty Without Trust

Renaud Foucart* Jonathan H.W. Tan†

September 26, 2019

Abstract

We propose and test a model of loyalty in games. Players can mutually develop loyalty by working towards a common goal that is Pareto-superior to any Nash equilibrium without it. Loyalty imposes a psychological cost on defecting in an ongoing cooperation, which is sufficient to sustain cooperation. Data on two dynamic games from a field experiment conducted in a Pakistani factory disentangle its effects from reciprocity and strongly support its validity. Loyalty begets cooperation, but unlike reciprocity it does not require individually costly trust. Loyalty holds across games and strengthens with social proximity.
Keywords: loyalty, cooperation, trust, reciprocity, field experiment

1 Introduction

The eventual success of a management team, joint venture, or strategic alliance in achieving its common goal relies on the mutual belief that members will cooperate and not shirk or defect midway for selfish gain. It is important for management scholars and practitioners to understand how such solidary pursuits we call *loyalty* emerge and persist. For example, repeated alliances between firms reduces strategic uncertainty through trust (Gulati, 1995). Trust is self-fulfilling if players believe that others reciprocate to repay kindness (Falk and

*Lancaster University Management School, r.foucart@lancaster.ac.uk

†Corresponding author: j.tan@ntu.edu.sg; Nanyang Technological University, Singapore. Acknowledgments: We are grateful to Bilal Riaz, Ali Riaz, Jamil Khan, Shamaz Khalid, Mustafa Zaman, and the assistants for the supporting the experiment. Thanks to Friedel Bolle, Yves Breitmoser, Paul Fenn, Alexander Kritikos, Yohanes Eko Riyanto and participants to the talk in Nottingham for the advice and encouragement.

Fischbacher, 2006) or avert guilt (Attanasi et al., 2016), or if players put a value on keeping their promises (Di Bartolomeo et al., 2019; Turmunkh et al., 2019). We, however, show that loyalty can yield cooperation even without costly investments of trust.

Loyalty is defined as a preference for upholding a mutually perceived cooperation. We propose a model where co-working towards a mutually beneficial common goal generates the belief that actions are consistent with this goal. Unlike trust, loyalty requires no individual cost in deviating from payoff maximization under self-interest, whereas treason yields disutility. This is sufficient to induce cooperation. We contribute to the economic and management sciences by distinguishing loyalty from trust by showing how loyalty emerges differently.

We use two dynamic games to disentangle loyalty from reciprocity (Falk and Fischbacher, 2006), guilt aversion (Attanasi et al., 2016), and other social preferences (Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Engelmann and Strobel, 2004) in a field experiment. We run it in a Pakistani textile factory for three main reasons. First, it is a natural setting where subjects are not biased by a training in game theory. Second, some subjects in this sample are already networked, letting us explore how trust and loyalty varies with social distance (Gulati, 1995; Glaeser et al., 2000). Third, we can incentivize subjects with high earnings worth up to a full day's wage.

The first game is an adaptation of the *Trust Game* (TG) of Berg et al. (1995). In Stage 1, the first mover chooses whether or not to send money to the co-player. This money is multiplied if sent. In Stage 2, the second mover chooses whether or not to return money to the first mover. By backward induction, under the assumption of pure self-interest the second mover is expected to abuse trust in Stage 2, so the first mover will not invest trust in Stage 1.

This equilibrium prediction is contradicted by the data. People send and return money due to trust and reciprocity. According to Falk and Fischbacher (2006), a second mover who is sufficiently motivated by reciprocity returns money to reciprocate the "kind" trust shown by the first mover in sending money. *Kindness* is defined as an expectation that the second mover gains more than the first from this process. Hence, there exists a reciprocal equilibrium in which the first mover is willing to trust if he expects a return with sufficiently high probability.

We compare behavior in the TG to that in a novel game where trust and reciprocity have no bite, and only loyalty can operate. Our game is inspired by the peer-lending microfinance arrangement commonly known as "committees" or Rotating Savings and Credit Associations

(Anderson and Baland, 2002, ROSCA). In Stage 1 of the *ROSCA game* (RG), two players simultaneously choose whether or not to join the committee and contribute to the common "pot." If anyone refuses, then the game ends. If both contribute, then Nature randomly selects one of the two players to win the pot. In Stage 2, the beneficiary from Stage 1 chooses whether or not to contribute to the pot again, which the co-player will now win.

In the RG, similar to self-interest, reciprocity predicts that both players contribute in Stage 1 but do not return anything in Stage 2. The reason is that contributing in Stage 1 signals neither expectations (Attanasi et al., 2016) nor kind trust (Falk and Fischbacher, 2006): by joining the committee a player *obtains* a higher expected payoff than by not joining and one that is exactly the same as the co-player's in any symmetric equilibrium. Thus, in Stage 2, one does not contribute again, because he has no expectations to meet so as to avert guilt, or have kindness to reciprocate. Fairness (Bolton and Ockenfels, 2000), non-intentions based reciprocity (Charness and Rabin, 2002), and efficiency concerns (Engelmann and Strobel, 2004) predict uniform behavior across the two games as their corresponding payoffs are similar.

In contrast, we observe that majority of subjects contribute in Stage 1 and return in Stage 2 in the RG. This behavioral pattern is compatible with our model of loyalty. When both players join the committee, their expected payoffs are increased. If the cost of betrayal is sufficiently high, then there exists an equilibrium in which both players contribute in Stage 1 and return in Stage 2. The other equilibrium however continues to exist: if a player believes the other would not return, then he will choose to not return. Stage 2 cooperation is therefore based on out-of-equilibrium beliefs of what the other would have done in that same role.

2 Theory

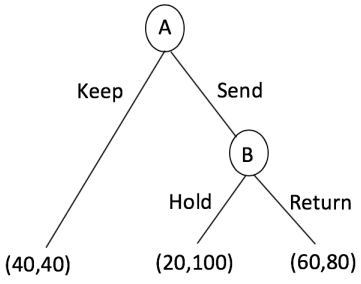
2.1 Trust Game

Our Trust Game (TG, see Figure 1a) is an adaptation of Berg et al. (1995). We discuss the basic game here and the experimental implementation in the next section. Two players A (he) and B (she) both receive an initial endowment y. In Stage 1, player A chooses to KEEP or SEND part of her endowment αy. If he chooses KEEP, the game ends and the payoffs are equal to the initial endowment. If he chooses SEND, the value of the amount sent increases and player B must choose to HOLD or RETURN part of the amount received. If she chooses

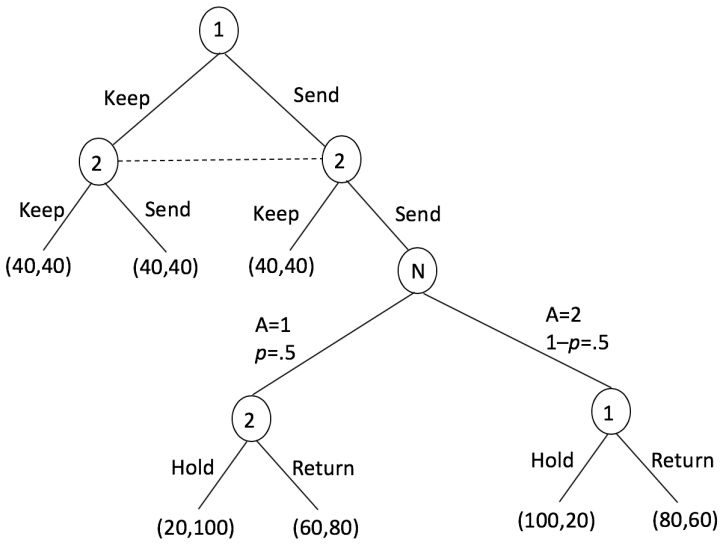
HOLD, their payoffs are $\{(1-\alpha)y, y+\beta_h y\}$, while if she chooses RETURN, their payoffs are $\{(1-\alpha)y+\frac{\beta_r}{2}y, y+\frac{\beta_r}{2}y\}$.

To preserve the properties of the original trust game, we make three restrictions on the parameters. First, with only self-interest, it is a dominant strategy for player B to HOLD ($\beta_h > \frac{\beta_r}{2}$). Second, if player B could commit to RETURN, player A would always prefer SEND ($\beta_r > 2\alpha$). Third, the total payoffs are higher when player A chooses SEND than when he chooses KEEP ($\beta_h > \alpha$). Assuming *self-interest*, the unique subgame perfect Nash equilibrium is (KEEP, HOLD).

Figure 1: Experimental Games, $y = 40$, $\alpha = \frac{1}{2}, \beta_h = 3, \beta_r = 4$



(a) Trust Game



(b) ROSCA Game

2.1.1 Reciprocity in TG

Falk and Fischbacher (2006) proposed a theory of *reciprocity* to explain observations of trust and reciprocity, which deviate from the self-interest prediction. A player B is reciprocal if she wants to return perceived kindness. In our version of TG, a player A choosing SEND in the first stage is always kind, as in doing so he expect the other player to receive a higher share of the payoffs. Hence, when the preference of player B for reciprocity (denoted ρ_B) is sufficiently high, player A can expect his kindness to be rewarded with strictly positive probability. We

provide a formal proof of this and Proposition 1 in the Appendix.

Proposition 1 *There exists a threshold reciprocity parameter $\tilde{\rho}_B$ such that, if $\rho_B \geq \tilde{\rho}_B$, the trust game displays a unique reciprocity equilibrium. Player A always chooses SEND in Stage 1, and player B chooses HOLD with a strictly positive probability $\tilde{p} < 1$ in Stage 2. Otherwise, if $\rho_B < \tilde{\rho}_B$, player A always chooses KEEP in the unique equilibrium of the game.*

For each value of ρ_B, there exists a probability p^* of HOLD, such that player B is indifferent between RETURN and HOLD. If p was higher than p^*, it means that player A is kinder than in equilibrium: he sends money while knowing it is unlikely to be returned. But then, player B would strictly prefer to RETURN to reciprocate on this kindness. Similarly, if p was lower than p^*, player B would strictly prefer to HOLD. For an equilibrium in which player A chooses SEND in the first stage to exist, he must therefore expect player B to RETURN with sufficiently high probability. This is the case if ρ_B is sufficiently high, as the equilibrium probability to HOLD p^* is decreasing in the importance of reciprocity for player B, ρ_B.

2.2 ROSCA Game

In Stage 1 of the ROSCA Game (RG, see Figure 1b), players 1 and 2 simultaneously choose to SEND or KEEP a share α of their endowment y. If at least one player chooses KEEP, similar to the TG the game ends and their monetary payoffs are $\{y, y\}$. If both choose SEND, Nature randomly selects one of the players with probability 1/2 to play a subgame similar to that of player B in the TG, while the other player has no further choices to make. Assume α, β_r and β_h are the same for both players, as well as the same parameter restrictions as TG.

Assuming only self-interest, there exists a subgame perfect Nash equilibrium in which both players choose SEND in Stage 1, and HOLD as player B. Indeed, in the last subgame, a player always receives a higher monetary payoff by choosing HOLD. By backward induction, player 1 expects that both he and player 2 HOLD with certainty when given a chance to do so. Hence, if a player believes the other would SEND, it is a best response to SEND if

$$\frac{1}{2}(1-\alpha)y + \frac{1}{2}(y + \beta_h y) > y, \tag{1}$$

which always holds as $\beta_h > \alpha$. An equilibrium in which both players choose KEEP in Stage 1 also exists, but it is not robust to a small tremble in the probability of the other player

choosing KEEP.[1]

2.2.1 Reciprocity in RG

The RG setting offers a twist that neutralizes the effect of reciprocity in a symmetric equilibrium. If both players expect each other to RETURN with the same probability, the expected payoffs are identical. Hence, the decision to SEND is neither kind nor unkind and in any symmetric equilibrium the utility (under reciprocity) is simply the monetary payoff. We provide a formal proof of this and Proposition 2 in the Appendix.

Proposition 2 *In the absence of loyalty concerns, the ROSCA game has a unique stable reciprocity equilibrium: both players choose SEND in Stage 1 and HOLD in Stage 2, for all values of the reciprocity parameters.*

In the reciprocity equilibrium of the RG, regardless of the reciprocity parameters, by symmetry players always choose HOLD as player B.[2] This leads to a "race-to-the-bottom" where players expect the other to HOLD and therefore feel no need to reciprocate. As there is no kindness involved, players choose SEND only to maximize expected payoffs. As in the case with only self-interest there is also another equilibrium in which everyone chooses KEEP, but it is not robust to trembles in the probability of the co-player choosing SEND.

2.3 Loyalty in RG

We define *loyalty* as a preference for maintaining a history of cooperation. We define cooperation as an action jointly taken by players to obtain a Pareto-improving outcome over noncooperation.[3] Our model assumes that loyalty does not derive any intrinsic benefit, but

[1] Even with an arbitrarily small probability $\epsilon > 0$ of player 1 choosing SEND, it is a best response for player 2 to also SEND.

[2] If you expect a player to RETURN less often than you, it is a best response for you to never RETURN.

[3] Denote the monetary payoff of player i at a final node n by $\pi_{i,n}$, the set of subgame perfect Nash equilibria of the game assuming only self-interest by J, and the node corresponding to one of those equilibria j by n_j^*, with $j \in J$. If there exists at least a final node n_k such that $\pi_{i,n_k} \geq \max_{j \in J} \pi_{i,n_j^*}$ for both $i \in \{1,2\}$ and $\pi_{i,n_k} > \max_{j \in J} \pi_{i,n_j^*}$ for at least one $i \in \{1,2\}$, denote by K the set of nodes satisfying the condition. Then two players cooperate if they make a joint decision under the beliefs that subsequent decisions will lead to a node $n_{k \in K}$ with certainty.

that betrayal incurs a cost given by the treason function $T(h)$, where $h = 1$ indicates a history of cooperation, otherwise $h = 0$. A necessary condition for treason is that the chosen action is incompatible with reaching the Pareto-improving outcome. For treason to happen however, there must be loyalty in the first place. This commitment is implicit and therefore weaker than promises (Di Bartolomeo et al., 2019). An action incompatible with reaching the Pareto-improving outcome thus yields a treason cost $T(h = 1) = \tau > 0$ and $T(h = 0) = 0$, depending of whether or not there is a history of cooperation, i.e. loyalty.

In RG, when two players have chosen to SEND in the first stage, the loyalty utility of player B when choosing to HOLD is thus given by

$$U_{B,h} = y + \beta_h y - (1 - p'_A)\tau, \tag{2}$$

where the last term represents the cost of treason. Denote the probability of player A choosing the strategy HOLD by p_A, the belief of player B on p_A by p'_A, and the belief of player A on p'_A by p''_A. By choosing HOLD, player B only betrays to the extent that she expects player A to have played cooperatively (i.e. with the intent of choosing RETURN in Stage 2 if in the counterfactual role). As she expects player A to choose RETURN with probability $1 - p'_A$, the expected cost of betraying him is $(1 - p'_A)\tau$. By choosing RETURN, player B receives $U_{B,r} = y + \frac{\beta_r}{2} y$, which is identical to the payoff under self-interest.

Proposition 3 *In the presence of loyalty concerns, there exists a threshold $\tilde{\tau}$ such that for sufficiently high cost of treason $\tau \geq \tilde{\tau}$, the ROSCA game displays two stable loyalty equilibria. In the first equilibrium, both players choose to SEND in Stage 1 and HOLD in Stage 2. In the second equilibrium, both choose SEND in Stage 1 and RETURN in Stage 2. If $\tau < \tilde{\tau}$, only the first equilibrium exists.*

The proof is in the Appendix. We start by identifying equilibria of the last subgame, assuming both players have chosen SEND. Consider a candidate symmetric equilibrium $p_A = p_B = p$. We immediately see that always choosing HOLD, i.e. $p = 1$, constitutes an equilibrium strategy in that subgame. If all players believe that no one cooperates, i.e. $p'_i = p''_i = 1$, $\forall i \in \{A, B\}$, then the best response of each player is to not cooperate as there is no treason cost involved.

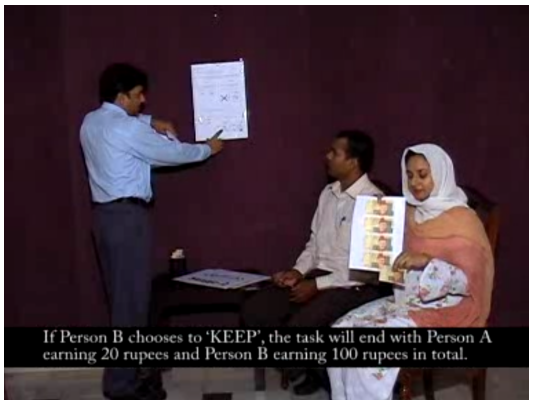
It is possible to identify a threshold treason cost $\tilde{\tau}$ such that, for all $\tau \geq \tilde{\tau}$, there exists two symmetric equilibria involving some cooperation in the last subgame. One is in pure strategy,

with $p = 0$ and both players always choosing RETURN. The other is a mixed strategy with $p = \tilde{p}$, which is however not robust to trembles in p as for all $p < \tilde{p}$ it is a best response to always RETURN (so that p decreases to 0) and for $p > \tilde{p}$ it is a best response to always HOLD (so that p increases to 1). As before, we can rule out asymmetric equilibria.

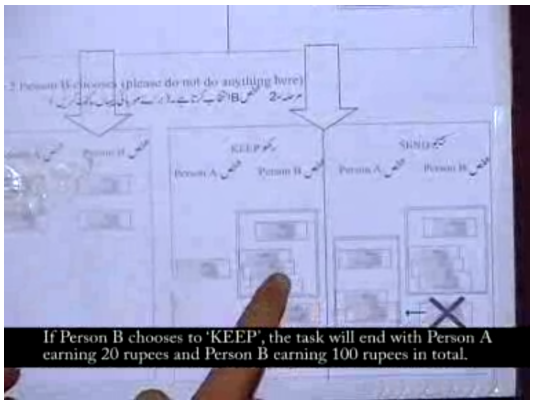
Inducing backwards to Stage 1, it is always a (weakly) dominant strategy for a player to SEND. If no one is expected to RETURN, then there is no loyalty and the result is similar to that under self-interest. If both players are expected to RETURN, then there is no betrayal in equilibrium and both players are better off to SEND. The expected loyalty utility at equilibrium is thus simply the monetary payoff.

3 Experiment

Figure 2: Experimental Instructions By Video



(a) Roles



(b) Closeup

The experiment was conducted in a textile factory in Lahore, Pakistan. Each session involved either TG or RG. We ran 4 sessions per game, each with 20 subjects, making it 160 subjects in total. The recruitment process was in itself an experimental manipulation to elicit social distance between subjects in the real world. We used sign up lists posted on a noticeboard, so people could see who had already signed up. We made it clear that "You can only participate in ONE session." We checked all sign up lists to ensure that no subject has signed up for more than one session, and if those who showed up were on the list.

Subjects received task booklets and payment sheets upon being seated. Three experimental assistants – none of whom were employees of the factory – administered each session. To maintain privacy, payments were calculated and administered by a separate person. To guarantee uniformity and quality in delivery across sessions, we instructed subjects using pre-recorded videos of the experimental instructions (see Fig. 2). In the video recorded instructions, one person acted as the experimenter, while the other two acted as Persons A and B. The video was scripted to complement the instructions with timed gestures and reference objects.

The video was filmed in spoken Urdu and English subtitles. Task booklets were presented in both Urdu and English. The video script and experimental material (available on request) was originally written in English, translated to Urdu, backtranslated to English by another person, and then compared to the original English version for consistency. Before subjects were allowed to start their games, we checked for understanding with a control questionnaire, and advised them verbally if answers were incorrect. After subjects had made their choices, we collected the task booklets and payment sheets and administered a post-experimental questionnaire. Subjects left after this was collected.

For the experimental implementations of TG and RG to be comparable, each subject played both roles A and B. Person B made decisions conditional on Person A choosing SEND. Playing both roles has been shown to induce empathy and more reciprocity (Burks et al., 2003), and this is a key feature of the RG. More importantly, it allows us to use TG to also test the robustness of mutual cooperation as a necessary condition for loyalty: if Person B gets to make a move, it means that Person A has trusted, and Person B will be loyal if she is *also* willing to SEND as Person A in the other task, and be more likely to RETURN. One's role was stated at the top left corner of the decision sheet. As "Person A," the person that one is matched with in that task is "Person B." Likewise, as Person B, the other person is Person A.

The games are as described above. In each stage, each person was endowed with an income of 20 rupees. In Stage 1 of the TG, Person A must choose between "KEEP" or "SEND" the 20 rupees. If Person A chose KEEP, then the task ended with each person earning 40 rupees in total. If Person A chose SEND, Person B received a coupon worth 80 rupees, by converting the 20 rupees she had and 20 rupees Person A sent.

In Stage 1 of the RG, both Person A and Person B chose whether to 'KEEP' or 'SEND' the 20 rupees. If one or both persons chose KEEP, then the task ended with each person

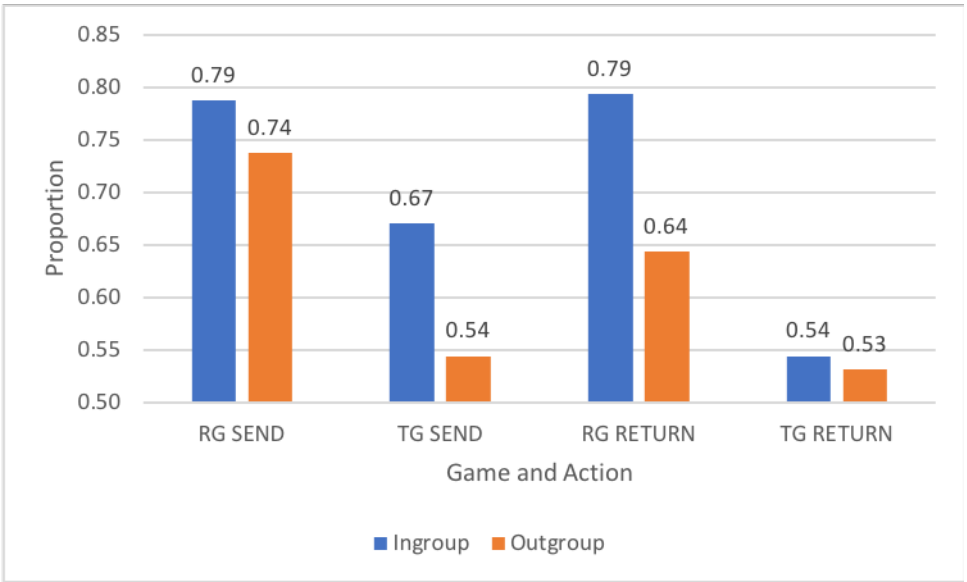
earning 40 rupees in total. If both persons chose SEND, Person B received a coupon worth 80 rupees, by converting the 20 rupees she had and 20 rupees Person A sent.

In Stage 2 of all games, Person B, who received the coupon in Stage 1, then chose to KEEP or SEND 20 rupees. Person B's choice was implemented only if Person A chose SEND in TG, or both chose SEND in RG.[4] If Person B chose KEEP, the task ended with Person A earning 20 rupees and Person B earning 100 rupees in total. If Person B chose SEND, Person A received a coupon worth 60 rupees in TG and RG, the task ended with Person A earning 60 rupees and Person B earning 80 rupees in total.

Subjects played one game with a randomly and anonymously matched co-player from the same session (ingroup), and another with someone from another session (outgroup). We expected social proximity to strengthen beliefs in cooperation and in turn trust and loyalty, but not reciprocity which is based on observed transfers. Subjects knew if the task involved the ingroup or outgroup. Therefore, subjects had four tasks in total, one for each of the two stages in each of the two games. One's payment was based on the choices that the matched pairs made in a "winning task," which was randomly determined by the spin of a wheel with four quadrants each representing a task. For the RG we also considered B's Stage 1 choice.

Social distance was varied within-subject and counterbalanced in alternating order: in the first and third sessions of each treatment subjects were matched with the ingroup then outgroup, and in opposite order for the second and fourth sessions. Pairwise matches were randomly determined by computer and printed out before the experiment, made available for inspection by independent monitors, and were used to calculate payments by the independent third party. The data was then entered into an excel spreadsheet and payments were calculated and put into envelopes and distributed. The expected payoff is 55.6 rupees, which was worth around half a day's wage.

Figure 3: Proportion of SEND and RETURN in TG and RG



4 Results

Figure 3 shows the probability of SEND and RETURN as a function of the game played.[5] Two-sample Wilcoxon rank-sum (Mann-Whitney) tests show that subjects SEND significantly more in RG than in TG to both the ingroup ($z = 1.65, p < 0.05$, 1-tail) and the outgroup ($z = 2.532, p < 0.01$). Subjects RETURN significantly more in RG than in TG (ingroup: $z = 3.094, p = 0.001$; outgroup: $z - 1.319, p < 0.1$). This result is compatible with the assumption of loyalty (proposition 3), but not with the assumption of only reciprocity (proposition 2).

Next, we compare behavior across matching protocols. Subjects SEND more to the ingroup (TG: $z = 1.543, p < 0.1$; RG: $z = 1.414, p < 0.1$), and RETURN more to the ingroup in RG ($z = 2.828, p < 0.01$), but not in TG ($n.s.$), i.e. social proximity strengthens trust and loyalty but not reciprocity. This is consistent with our theory that reciprocity is based on observed transfers, while trust and loyalty are based on beliefs, which can be reinforced by social proximity.

[4]For the experiment, we kept the terms 'KEEP' and 'SEND' symmetric across both roles. In the other sections of this paper, we distinguish Person B's actions from A's by labelling them 'HOLD' and 'RETURN'.

[5]The means for RG exclude subjects who chose KEEP in Stage 1, as choosing KEEP ends the game and prevents them from proceeding to Stage 2.

Finally, to test the validity of loyalty across games, we match each subject's choices in both roles of TG. If there is loyalty, Person B should be loyal if she is also willing to SEND as Person A in the other task. Therefore, we check if those who chose to SEND will RETURN more frequently. When matched with the ingroup, those who chose SEND as player A chose RETURN 64.2% of the time, while those who chose KEEP as Person A RETURN 34.5% of the time, and this difference is statistically significant ($z = -2.461, p < 0.01$). This result is robust to matching with the outgroup, with those who SEND choose RETURN 67.4%, versus 36.1% for those who KEEP ($z = -2.762, p < 0.01$).

5 Conclusion

Self-interest, reciprocity, and loyalty predict different behavioral patterns for TG and RG. This allows a clear identification of motives. Reciprocal players are predicted to RETURN in Stage 2 of the TG if the reciprocity parameter is sufficiently high, but like self-interest, they HOLD in Stage 2 of the RG, but loyalty predicts RETURN. We observe a majority of subjects choosing RETURN in RG, and significantly more than in TG. Loyalty explains successful business ventures motivated by prospective profit, which suffices to overcome transaction costs in the sense of Gulati (1995) even in the absence of kind trust. This further relaxes the assumptions of social preferences in explaining business transactions.

The loyalty equilibrium relies on beliefs over strategies that are never played. One co-operates in Stage 2 only because one believes the co-player also intended to cooperate in this stage. While these beliefs cannot be confirmed without repeated play, more cooperation can be expected in the ingroup. Indeed, we found increased loyalty (and trust but not reciprocity) in ingroup interactions. To corroborate the role of loyalty beliefs, those who chose SEND as Person A in the TG were almost twice as likely to RETURN as Person B. Our evidence of loyalty in games as a distinct preference warrants further research.

References

Anderson, S. and Baland, J.-M. (2002). The economics of roscas and intrahousehold resource allocation. *The Quarterly Journal of Economics*, 117(3):963–995.

Attanasi, G., Battigalli, P., and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193.

Burks, S. V., Carpenter, J. P., and Verhoogen, E. (2003). Playing both roles in the trust game. *Journal of Economic Behavior & Organization*, 51(2):195–216.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2019). Promises, expectations & causation. *Games and Economic Behavior*, 113:137–146.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4):857–869.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., and Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics*, 115(3):811–846.

Gulati, R. (1995). Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Academy of management journal*, 38(1):85–112.

Turmunkh, U., van den Assem, M. J., and Van Dolder, D. (2019). Malleable lies: Communication and cooperation in a high stakes tv game show. *Management Science*.

Appendix

A1: Reciprocity in the trust game

First we consider beliefs in Stage 2. Denote the probability of player B choosing the strategy HOLD by p, the belief of player A on p by p', and the belief of player B on p' by p''. Player B's payoff from playing HOLD is given by

$$U_{B,h} = y + \beta_h y + \rho_B \times p'' \beta_h y \times (1 - p'')(-\frac{\beta_r}{2} y). \tag{3}$$

The material payoff from HOLD, $y+\beta_h y$, enters utility linearly. The reciprocity parameter ρ_B is a positive constant capturing the strength of player B's reciprocal preferences. Player A's "kindness" is given by $p''\beta_h y$. When player A chooses SEND, he effectively commits to a weakly higher share to player B, with the belief that with probability p' player B will be "unfair" and HOLD, and with complementary probability player B will be "fair" and choose RETURN.[6] As B does not know with certainty A's expectation of her to RETURN, B's evaluation of player A's kindness is conditioned on p'', i.e. player B's belief of player A's expectation of what she will choose. The third term, $(1 - p'')(-\frac{\beta_r}{2} y)$ is the reciprocation term. HOLD gives player A, who was hoping for an additional $\frac{\beta_r}{2} y$ with probability $(1 - p')$, less than expected. Player B therefore suffers from choosing HOLD and not reciprocating A's kindness.

Player B's payoff from playing RETURN is given by

$$U_{B,r} = y + \frac{\beta_r}{2} y + \rho_B \times p'' \beta_h y \times p''(\frac{\beta_r}{2} y). \tag{4}$$

By choosing RETURN, B's monetary payoff is decreased to $y+\frac{\beta_r}{2} y$. The reciprocity parameter and the kindness term are as before, as the strength of reciprocity is exogenously given, and kindness depends on what player B believes player A expects her to choose. Player B now gives a higher than expected payoff to player A, as he was expecting player B to KEEP with probability p'. Again, player B uses her beliefs p'' of what player A expects her to do as she does not know p'.

We then turn to the choice of player A in Stage 1. In equilibrium, $p' = p'' = p$ must hold. Moreover, as the amount sent is discrete, player A cannot influence p by marginally increasing

[6] Falk and Fischbacher (2006) define "fairness" using an equal sharing rule, which is obtained by (SEND, RETURN).

α. Thus, the utility from reciprocating is equal to *zero* in expectation. The payoff of player A, assuming equilibrium in the last subgame, is therefore entirely determined by her expected monetary payoff

$$\begin{aligned} U_{A,k} &= y, \\ U_{A,s} &= (1-\alpha)y + (1-p')\frac{\beta_r}{2}. \end{aligned} \tag{5}$$

A2: Proof of Proposition 1:

Proof. Stage 2: The difference between the payoffs of the two strategies simplifies to

$$U_{B,r} - U_{B,h} = \frac{y}{2}(\beta_r - 2\beta_h + p\rho_2\beta_r\beta_h y). \tag{6}$$

It is never a best response in this subgame to always return, as for $p = 0$ player B always prefers to hold, $U_{B,r} < U_{B,h}$. The reason is that if player B believes that player A expects her to always return, there is no kindness in player A choosing to send, and therefore no benefit to reciprocate. At equilibrium, if p has an interior solution, we are looking for a value of $p = p' = p''$ such that $U_{B,r} - U_{B,h} = 0$. This solution is given by

$$\tilde{p} = \frac{2\beta_h - \beta_r}{\beta_h\beta_r\rho_B y}, \tag{7}$$

so that the equilibrium probability of player B Keeping the investment is $p = \min\{\tilde{p}, 1\}$, with $p = 1$ for $\rho_B \geq \frac{2\beta_h - \beta_r}{\beta_h\beta_r y}$. Note that as $\tilde{p} > 0$ for all ρ_B, there is no corner solution in which player 2 returns the investment all the time. The reason is that if player A expects player B to return the investment all the time, there is no kindness in Sending in the first stage, so no incentive to reciprocate. The mixed strategy equilibrium of this last subgame is "stable" in the sense that the best response to any small perturbation to the equilibrium p'' would bring this probability back to equilibrium: If p'' is strictly higher than the equilibrium value, player B strictly prefers to return, and if it is strictly lower player B strictly prefers to keep.

Stage 1: Turning now to the choice of player A, his payoff is given by:

$$u_{A,k} = y \tag{8}$$

$$u_{A,s} = (1-\alpha)y + (1-p')\frac{\beta_r y}{2} + \rho_A(p - \frac{\beta_r y}{2})(p''\frac{\beta_r y}{2} - p\frac{\beta_r y}{2}) \tag{9}$$

At equilibrium in the last subgame, it must hold that $p' = p'' = p$, so that (9) rewrites:

$$u_{A,s} = (1-\alpha)y + (1-p)\frac{\beta_r y}{2}. \tag{10}$$

Hence, player A prefers to Send if the expected amount he receives back $(1-p)\frac{\beta_r y}{2}$ is higher than the amount sent αy. The condition for A to Send in the first stage is thus given by

$$p \leq 1 - 2\frac{\alpha}{\beta_r} = \bar{p}. \tag{11}$$

With, by assumption, $\beta_r > 2\alpha$, so that the maximum probability of player B keeping the amount with player A still sending is $\bar{p} \in [0,1]$. Using the expression of $\tilde{p}$ found in 7 we can express the condition $p \leq \bar{p}$ in terms of the reciprocity parameter of player B, we find

$$\rho_B \geq \frac{2\beta_h - \beta_r}{\beta_h(\beta_r - 2\alpha)y}. \tag{12}$$

■

A3: Reciprocity in the ROSCA game

Let us analyze the RG under the assumption of reciprocity *à la* Falk and Fischbacher (2006). Without loss of generality, we begin with the subgame of the player selected by Nature, and call this player B as we did for TG. Denote the probability of each player choosing KEEP conditional on being assigned the respective roles by Nature by p_A and p_B. Denote the belief of player j on p_i as p_i', and the belief of player i on p_i' as p_i''.

Player B's payoff from playing HOLD is given by

$$U_{B,h} = y + \beta_h y + \rho_B \nu(\frac{1}{2}p_B''\beta_h y - \frac{1}{2}p_A'\beta_h y)(1-p_B'')(-\frac{\beta_r}{2}y). \tag{13}$$

There are two differences with the analysis of TG. First, we must consider "intentions" (Falk and Fischbacher, 2006) $\nu(p', p'')$, with $\nu = 1$ corresponding to a fully intentional action and $\nu < 1$ if otherwise. In contrast to TG, the intentions of a player choosing SEND are now ambiguous. A player who has been assigned the role of A might be expecting the co-player to RETURN, despite intending to HOLD as player B in the counterfactual role, and therefore was not fully intending to be kind. Second, the kindness term $(\frac{1}{2}p_B''\beta_h y - \frac{1}{2}p_A'\beta_h y)$ depends on the difference between the expected strategy of both players in the role of player B. A player's action is considered kind if he chooses RETURN with a higher probability than the co-player.

Similarly, player B's payoff from playing RETURN is given by

$$U_{B,r} = y + \frac{\beta_r y}{2} + \rho_B \nu(\frac{1}{2}p''_B\beta_h y - \frac{1}{2}p'_A\beta_h y)p''_B(\frac{\beta_r}{2}y). \tag{14}$$

A4: Proof of Proposition 2:

Proof. Stage 2: Consider the possible equilibria. We start by looking at the symmetric case $p_A = p_B$. At equilibrium, we can rewrite (13) and (14) as

$$U_{B,h} = y + \beta_h y \tag{15}$$

$$U_{B,r} = y + \frac{\beta_r y}{2}, \tag{16}$$

and it immediately follows that as $\frac{\beta_r}{2}y < \beta_h y$, $U_{B,h} > U_{B,r}$. Hence, in any symmetric equilibrium it is a best response for both players to choose HOLD when they are selected in Stage 2. Now consider an asymmetric case, $p_A > p_B$. This assumption implies $p_A > 0$. However, for $p_A > p_B$, the kindness term is strictly negative for player B, so that

$$U_{B,h} = y + \beta_h y + \rho_B \nu(\frac{1}{2}p''_B\beta_h y - \frac{1}{2}p'_A\beta_h y)(1 - p''_B)(-\frac{\beta_r}{2}y) > y + \beta_h y \tag{17}$$

$$U_{B,r} = y + \frac{\beta_r y}{2} + \rho_B \nu(\frac{1}{2}p''_B\beta_h y - \frac{1}{2}p'_A\beta_h y)p''_B(\frac{\beta_r}{2}y) < y + \frac{\beta_r y}{2}. \tag{18}$$

Hence, it is a best response for B to always HOLD, $p_B = 1$. This contradicts the assumption that $p_A > p_B$.

Stage 1: We know from Stage 2 that in any equilibrium it must hold that $p = p' = p'' = 1$. If both players choose SEND, each player receives on expectation $y + \frac{1}{2}(\beta_h - \alpha)$ and, if at least one player chooses KEEP, each receive y. Hence, if there is a strictly positive probability of the co-player choosing SEND, it is a best response for a party to also SEND. ■

A5: Proof of Proposition 3

Proof. Stage 2: We start by expressing the condition $U_{B,r} \geq U_{B,h}$ – whether player B prefers RETURN in a symmetric equilibrium - as a condition on the equilibrium probability of both players choosing HOLD p:

$$p \leq 1 - y\frac{2\beta_h - \beta_r}{2\tau} = p^f. \tag{19}$$

Thus, for τ sufficiently high, $p_f \geq 0$,

$$\tau \geq \frac{2\beta_h - \beta_r}{2} y = \tilde{\tau}. \tag{20}$$

Stage 1: Going back to Stage 1 and given a symmetric equilibrium p in the last subgame, player A (symmetric for player B) prefers SEND for any strictly positive probability of player A choosing SEND if and only if

$$\frac{1}{2}\left(p(y + \beta_h y - (1-p)\tau) + (1-p)(y + \frac{\beta_r}{2} y)\right) + \frac{1}{2}\left(p(y - \alpha y - \tau) + (1-p)(y - \alpha y + \frac{\beta_r}{2} y)\right) \geq y. \tag{21}$$

There is therefore no additional condition on the cost of treason, other than the ones made for the last subgame, to consider when studying the possibility of an equilibrium in which both players SEND and RETURN with probability 1. Both players (selfishly) choosing SEND in Stage 1 but choosing HOLD in Stage 2 remains a subgame perfect Nash equilibrium as if there is no cooperation expected ($p = 1$). For the pure strategy equilibrium with $p = 0$, no one is expected to betray in Stage 2 if and only if $\tau \geq \tilde{\tau}$, so that the expected cost of betraying and being betrayed are both equal to zero, so SEND is again a best response in Stage 1. ■