IN PRESS, Journal of Economic Behavior and Organization

POSITIVE RECIPROCITY AND INTENTIONS IN TRUST GAMES

KEVIN A. MCCABE, MARY L. RIGDON* AND VERNON L. SMITH

Abstract. Several recent theories in behavioral game theory aim at explaining the behavior of subjects in experimental bargaining games. These models can be partitioned into two classes: outcome-based and intention-based. Outcome-based models treat the intentions that players attribute to one another as unnecessary for predicting behavior. Intention-based approaches, and in particular the trust and reciprocity hypothesis, rely on this attribution of intentions in an essential way. We report laboratory data from simple two-person trust games which is inconsistent with outcome-based models, but predicted by the trust and reciprocity hypothesis.

JEL Classification: C72, C78, C91

1. Introduction

In two-person exchange whoever moves first may give up a sure-thing with a certain value in exchange for an anticipated future benefit. Receiving the future benefit, however, is contingent on how the second mover reacts to the first mover's decision. Intuitively, the second mover can either pursue her dominant action (which may leave the first mover with a loss) or reciprocate to achieve a joint maximum to be shared by both movers. Each, therefore, incurs an opportunity cost to arrive at the joint benefit. There are many examples of two-person exchange environments. A sister lets her younger brother go first in a computer game with the understanding that she will get a longer turn later. A couple might go to a Cubs' game one evening with the understanding that the next week they will attend a play. A buyer on the Internet buys a good—sight unseen—only to receive the goods in a later shipment. An example familiar from labor economics is when a firm offers an employee a wage above the market-clearing level, expecting

Date: March 22, 2002.

Corresponding author. We would like to thank Rachel Croson, Martin Dufwenberg, Anthony Gillies, Daniel Houser, and the participants at the Economic Science Association meetings in Tucson (October 2000) and in Amsterdam (October 2000) for discussion and comments.

that in exchange the worker will provide greater effort (thus achieving a cooperative outcome). We will model such environments by *two-person trust games*.

There is ample experimental evidence that suggests a considerable proportion of play in two-person trust games deviates from that predicted by standard non-cooperative game theory (Berg,*et al.*, 1995; McCabe,*et al.*, 1998). A significant percentage of anonymously paired subjects arrive at cooperative outcomes. There are two classes of models that attempt to explain these results (as well as the observed behavior in a variety of experimental games). One approach focuses exclusively on properties of the outcomes in these games. For example, models which posit that a certain proportion of the population is altruistic or spiteful (Levine, 1998), or have certain thresholds of inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), all fall within the class of outcome-based models. A second approach emphasizes the role of intentions in achieving cooperative outcomes in personal exchange. The models in McCabe and Smith (2000), Dufwenberg and Kirchsteiger (1998), and Falk and Fischbacher (1998), for example, fall within the class of intention-based accounts. Whereas the outcome-based approaches imply that intentions are superfluous, intention-based models rely essentially on players reading each other's motives (and not merely their actions).

One consequence of the intention-based approach is that depending on the available alternatives, identical outcomes may be interpreted differentially. For outcome-based approaches this is not the case. Only the intrinsic properties of outcomes are assumed to drive behavior so what alternatives the players face is irrelevant. In order to test between these two approaches, we design a treatment variable that varies Player 1's opportunity cost between zero (in the involuntary trust game) and positive (in the voluntary trust game). According to an intention-based approach (and in particular the trust and reciprocity hypothesis), Player 2 must consider the motives of Player 1. We hypothesize that this mind-reading is a function of Player 1's opportunity cost. Therefore, these approaches predict that the cooperative move by Player 1 in the positive opportunity cost games will generate greater reciprocity from Player 2 than the same move in the zero opportunity cost game. While such results are consistent with the TR hypothesis, we will see that they are inconsistent with the behavior predicted by outcome-based models.

The paper is organized as follows. The next section (Section 2) has a more detailed discussion of the trust and reciprocity hypothesis. Section 3 provides an overview of two recent outcome-based models. Section 4 contains the two experimental treatments and Section 5 contains the predictions and hypotheses for our design. Lastly, Section 6 reports the experimental protocol and the results.

2. Trust and Reciprocity

Within the class of intention-based approaches, we want to focus on the trust and reciprocity (TR) hypothesis and how intuitively it can explain deviations from standard non-cooperative theory observed in laboratory experiments with two-person trust games. The deviations are two-fold. First, in trust games, for Player 1 to achieve a future benefit, he must deviate from the subgame perfect strategy profile in the game. Second, a significant portion of Players 2 (positively) reciprocate instead of playing their dominant strategies. Positive reciprocity can be described as the costly behavior of a second mover that rewards a first mover based on both the gains from exchange to the second mover as well as the second mover's beliefs about the intentions motivating the action of the first mover.[1]

The TR hypothesis explains this behavior as a *reciprocal-trust relationship* between Players 1 and 2. Player 1 and Player 2 are reciprocally-trust related if (i) there are mutual gains from their joint actions, (ii) Player 1 takes a risk by trusting Player 2, and (iii) Player 2 gives up something in order to reciprocate Player 1's trust. The mutual gains from the exchange are measured relative to the subgame perfect equilibrium (SPE). So the first condition ensures that if Players 1 and 2 are in a reciprocal-trust relationship, they will reach an outcome which is Pareto superior to that prescribed by non-cooperative game theory. The second condition brings Player 1's opportunity cost into the relationship in an explicit way. If Player 1's sure-thing option is zero, then there is little risk in his opting to try for another outcome. If Player 1's opportunity cost is positive, then taking the risk to achieve a cooperative outcome can signal Player 1's intentions toward Player 2—namely, the intention to enter into a reciprocal-trust relationship. Finally, in order for Player 2 to reciprocate, she must not be playing her dominant strategy.

[1]By way of contrast, *negative* reciprocity is essentially a punishment strategy, in which one party incurs a cost to punish another for failing to reciprocate.

Notice that a reciprocal-trust relationship is not merely identified with a profile of actions. Consider condition (ii). Player 1 trusts Player 2 only if Player 1 has two relevant *beliefs*: that Player 2 will interpret his move as a trusting one, and that Player 2 will reciprocate. And, as for condition (iii), it is clear that Player 2's action can be described as reciprocal only if she *interprets* Player 1's action as trusting. That is, Player 2 must attribute to Player 1 the *intention* of entering into a reciprocal-trust relationship.

Such an attribution of intentional states to others is part of what cognitive scientists call *mentalizing* or *folk psychology* (Baron-Cohen, 1995). Humans routinely explain the behavior of others by attributing to them mental states of various sorts: beliefs, desires, and so on. Likewise, given some attribution of mental states, we unconciously and routinely predict how others will behave. According to Simon Baron-Cohen, subjects must have a shared attention on possible mutual gains. It is not enough for Player 2 to infer "Player 1 moved down because 1 wants more money." Instead, Player 2 must infer on the basis of Player 1's action "1 moved down because 1 sees that 2 sees this as a reciprocal-trust relationship." The TR hypothesis therefore suggests that Player 2 can read the action of Player 1 as signalling trust that Player 2 will reciprocate if given the chance. Player 1, knowing that this signal can be interpreted by Player 2, reduces his assessed risk in forgoing the sure-thing. Under the TR hypothesis, it follows that the formation of the second mover's beliefs about the intentions of the first mover must be understood to include the opportunity cost of the first mover's action.

3. Outcome-based Models

Here we briefly outline two recent outcome-based models: ERC (Bolton and Ockenfels, 2000) and the Fehr-Schmidt model (1999). In Section 5 we will derive specific predictions for our treatments.

Bolton and Ockenfels propose in their ERC preference model for two-person games a motivation function (which : $v_i = v_i(y_i, \sigma_i)$ where y_i is i's own payoff and $\sigma_i = \frac{y_i}{y_1+y_2}$ for $i = 1, 2$. So the motivation function depends on Player i's own monetary payoff and the relative share of the payoff that i is receiving. There is a tradeoff between how much agents value their own payoff and their relative share of the total payoff in an outcome. The ERC model types players according to where

these thresholds occur. For our purposes it is enough to note that the thresholds are solely functions of intrinsic properties of outcomes: namely, i's own monetary payoff and the distribution of the total payoff.

Fehr and Schmidt also propose a model based on inequity aversion. Again, we restrict ourselves to the special case of two-person games. Let $\vec{x} = \{x_1, x_2\}$ be the vector of payoffs to Players 1 and 2 for a given outcome. Player i's Fehr-Schmidt utility function is: $U_i(\vec{x}) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\}$, where $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$. In this model, α_i measures how much Player i dislikes inequitable outcomes which favor Player j and β_i measures how much Player i dislikes inequitable outcomes which favor himself. These two measures determine player types in the population (which are assumed to be uniformly distributed). Again it is only intrinsic properties of outcomes that can be used in this model to explain behavior.

4. Experimental Treatments

We consider our two treatments: the voluntary trust game (VTG) and the involuntary trust game (ITG).

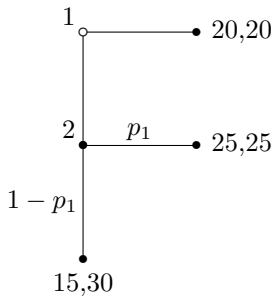

Figure 1. Voluntary trust game (VTG)

The voluntary trust game is represented in Figure 1. Player 1 has an outside option of $[20, 20]$ which is the SPE. If Player 1 moves down, Player 2 has a choice between the symmetric joint maximum outcome of $[25, 25]$ or the defection outcome of $[15, 30]$.

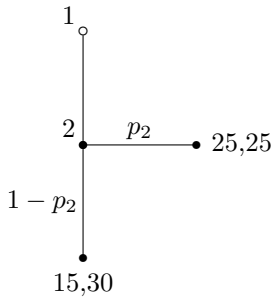

FIGURE 2. Involuntary trust game (ITG)

Compare Figure 1 to the involuntary trust game in Figure 2. The only difference between the two games is that Player 1 does not have an outside option in the involuntary trust game. The treatment variable varies the opportunity cost for Player 1 between positive (in the VTG) and zero (in the ITG).

5. Predictions and Hypotheses

The behavior of interest in these treatments is the relative rates of cooperation by Players 2 (i.e., comparing p_1 and p_2). It is straightforward to see that outcome-based approaches—and in particular ERC and Fehr-Schmidt—all predict that cooperation rates of Players 2 should not vary across the voluntary and involuntary trust treatments.

The ERC prediction is as follows. After a move down by Player 1, Players 2 in the voluntary and involuntary games have identical choices available to them. Therefore, the probability of a right move by Player 2 is the same in both games. That is, $p_1 = p_2$.[2]

For the Fehr-Schmidt prediction we need only look at the adjusted utilities of second movers. In the VTG, for the cooperative $[25, 25]$ outcome, Player 2's utility is 25. In the ITG, Player 2 has exactly the same choices and possible outcomes, and so the same adjusted utility of 25 at this outcome. The utilities for the defection outcome $[15, 30]$ are also identical across these games. This is because the value of

[2]This prediction in fact has the same form as the ERC prediction in the mini-best shot and mini-ultimatum games—see the proof of Statement 7 (p.176) of Bolton and Ockenfels.

β_2 is assumed to be uniformly distributed. It follows, then, that the Fehr-Schmidt model predicts the probability of cooperation by Player 2 will be the same in both treatments: $p_1 = p_2$.[3].

The TR hypothesis offers a very different prediction across these treatments. In the ITG we remove Player 1's ability to send cooperative signals to Player 2 by eliminating Player 1's opportunity cost to trust. The result is that from Player 2's perspective, there is no longer an ability to read the intentions of her counterpart. According to the TR hypothesis, this should significantly reduce the amount of cooperative play. Such conditions significantly reduce Player 2's ability to reciprocate because she cannot reliably attribute intentions of trust to Player 1. Therefore, in the ITG, we should observe more play at the $[15, 30]$ outcome.

The TR hypothesis predicts that the cooperative move in the positive opportunity cost games will generate greater reciprocity than the same move in the zero opportunity cost game. That is, TR predicts that p_1 will be significantly greater than p_2.

$$\begin{aligned} H_0 : p_1 - p_2 &\leq 0 \\ H_1 : p_1 - p_2 &> 0 \end{aligned}$$

where, as before, p_1 is the proportion of moves at $[25, 25]$ conditional on Player 1's move down in the VTG, and p_2 is the proportion of moves at $[25, 25]$ conditional on Player 1's move down in the ITG. The predictions of outcome-based models are represented under our null hypothesis.

6. Procedures and Results

In all experiments, subjects were paid \$5.00 for arriving on time. At the end, their accumulated earnings were paid to them privately (single-blind protocol). The interactions consisted of anonymous and random pairings in a one-shot computerized game. The payoffs are actual (US) dollar amounts the subjects could earn, and are common information. The subjects were undergraduates at the University

[3]An alternative version of the ITG would be to have Player 1's outside option be $[0, 0]$. While this is an interesting empirical variation, the predictions of outcome-based models about the level of cooperative play remain unchanged.

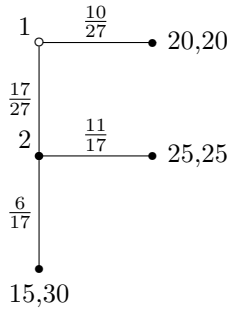

FIGURE 3. Frequency of moves in the Voluntary trust game

of Arizona and did not have prior experience with a trust environment. Each experimental session consists of 12 subjects, six assigned to each treatment condition. As a subject enters the Economic Science Laboratory, he is paid his show-up fee and immediately seated at a computer terminal in a large room containing 40 terminals. Each terminal is in a separate stall, and the 12 subjects are dispersed so that no subject can see the terminal screen of another. Each is randomly assigned to one of the treatments, then to one of six pairs, and finally randomly assigned a role (Player 1 or 2). The game is played sequentially. The experiments lasted on average 30 minutes, from arrival to completion. Each subject participates in one and only one of the treatments.

Figure 3 records the proportion of observed play at each node in the VTG, and Figure 4 records the proportion of observed play at each node in the ITG. The null hypothesis that the proportion of cooperative outcomes is identical across treatments is easily rejected by both a t-test and a bootstrap test ($p < 0.01$). There is a significant treatment effect between the two environments.

7. Discussion

The data in these simple experiments are inconsistent with the predictions of the ERC and Fehr-Schmidt models. What is instructive is that all of these models predict the same behavior—and for largely the same reasons—in the voluntary and involuntary trust games and this should cast doubt on outcome-based explanations in general. On the other hand, it is consistent with—indeed, predicted by—the TR

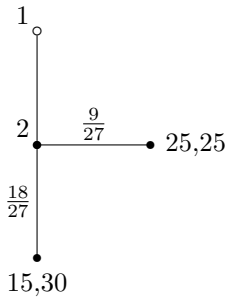

FIGURE 4. Frequency of moves in the Involuntary trust game

hypothesis that cooperative play occurs significantly less often in the involuntary trust game. By eliminating Player 1's opportunity cost associated with playing down, we have restricted Player 2's ability to unambiguously read her counterpart's intentions. In the voluntary trust game, an intentional move down the tree by Player 1 can be explained by Player 2 as an act of trust. In the involuntary game, however, a down move carries no such information because Player 1 had no choice but to move down in the game.

The data reported here are not the only data that inequity aversion models have trouble explaining. One such experimental treatment is the single- versus double-blind protocol in dictator games. The outcome-based utilities are the same across treatments, but the results are very different. The number of self-interested offers with the double-blind protocol is much larger than with single-blind payoffs (Hoffman,*et al.*, 1996). Another procedural effect that outcome-based approaches are unable to explain is how alternative descriptions of a player's counterpart in the instructions can impact behavior in bargaining environments. Systematically referring to one's counterpart as a "partner" in one treatment, and as an "opponent" in the other, changes observed behavior in an extensive form trust game (Burnham,*et al.*, 2000). Again, the adjusted utilities across treatments do not vary, and so outcome-based models predict no difference.

A similar conclusion about the inadequacy of outcome-based models is reached by Falk,*et al.* (2002). However, there are significant differences between the two studies. First, they make use of mini-ultimatum games and not trust games. Pure

trust games allow us to isolate opportunity costs, and vary this without introducing negative reciprocity. Second, their design has a serious confounding factor: each second mover indicates her action at *both* decision nodes (for the case of a left branch offer and for the case of a right branch offer) without knowing what the first mover has actually proposed. The games, therefore, are played in strategic (or normal) form. It is well documented that the extensive and strategic forms are played differently (McCabe,*et al.*, 2000; Schotter,*et al.*, 1996).[4] Furthermore, psychological literature suggests that when second movers are asked to make hypothetical decisions, like "What would I do if Player 1 moves left?" and "What would I do if Player 1 moves right?", what subjects report they would choose can be very different from what they actually choose (Langer, 1975).

An interesting variation to the experiments discussed here would be to hide the value of Player 1's opportunity cost (i.e. the value to Player 1 of his outside option) from Player 2. Under the TR hypothesis, anything that makes the signal to Player 2 about Player 1's intentions more noisy should reduce the likelihood of observing cooperation. By having payoff privacy at the outside option node, it will be common information that Player 1 actually has a choice to make, but it is unclear to Player 2 whether Player 1 is taking a risk to achieve the cooperative outcome or playing a weakly dominant strategy. TR would predict that conditional on Player 2 having a move, play in this game would not look significantly different from that observed in the involuntary treatment. However, we hypothesize that in such noisy environments Players 1 predict that their intentions will not accurately be read. And so TR would predict more play at the SPE. Further experiments need also to test the boundary conditions for reliable intentionality detection in two-person trust games.

[4]In fact, we think that these types of decision making environments may be what is driving some of the different results found in Dufwenberg and Gneezy (2000) and Charness and Rabin (2001). In order for such results to go through one needs the auxiliary hypothesis that responders in these games behave the same regardless of whether they *actually* see the first mover's choice or are just told to *assume* that the first mover has chosen a particular action. So the experiments reported here offer a more direct test. Interestingly, Nelson (2002) employs the strategy method in a truncated ultimatum game and finds the data are consistent with an intention-based approach.

References

Baron-Cohen, S., (1995). Mindblindness. Cambridge, MA: MIT Press.

Berg, J., Dickhaut, J., McCabe. K., 1995. Trust, Reciprocity, and Social History. Games and Economic Behavior 10, 122–142.

Bolton, G., Ockenfels, A., 2000. ERC: A Theory of Equity, Reciprocity, and Competition. American Economic Review 90(1), 166–193.

Burnham, T., McCabe, K. Smith, V., 2000. Friend-or-foe Intentionality Priming in an Extensive Form Trust Game. Journal of Economic Behavior and Organization 1244, 1–17.

Charness, G., Rabin, M., 2001. Understanding Social Preferences with Simple Tests, Working Paper, Univ. Pompeu Fabra, No.441.

Dufwenberg, M., Gneezy, M., 2000. Measuring Beliefs in an Experimental Lost Wallet Game. Games and Economic Behavior 30(2), 163–182.

Dufwenberg, M., Kirchsteiger, G., 1998. A Theory of Sequential Reciprocity, Tilburg CentER for Economic Research Discussion Paper: 9837.

Falk, A., and U. Fischbacher (1999). "A Theory of Reciprocity," Institute for Empirical Research in Economics, The University of Zurich, Working Paper No.6.

Falk, A., Fehr, E., Fischbacher, U., 2002. On the Nature of Fair Behavior. Economic Inquiry, forthcoming.

Fehr, E., Schmidt, K. M., 1999. A Theory of Fairness, Competition, and Cooperation. Quarterly Journal of Economics 114(3), 817–868.

Hoffman, E., McCabe, K., Smith, V., 1996. Social Distance and Other-Regarding Behavior in Dictator Games. American Economic Review 86, 653–660.

Kagel, J. H., Wolfe, J., 2001. Tests of Difference Aversion to Explain Anomalies in Simple Bargaining Games. Journal of Experimental Economics 4, 203–219.

Langer, E., 1975. The Illusion of Control. Journal of Personality and Social Psychology 32, 311–328.

Levine, D. K., 1998. Modeling Altruism and Spitefulness in Experiments. Review of Economic Dynamics 1(3), 593–622.

McCabe, K., Rassenti, S. and Smith, V., 1998. Reciprocity, Trust, and Payoff Privacy in Extensive Form Bargaining. Games and Economic Behavior 24, 10–24.

McCabe, K., Smith, V., 2000. Goodwill Accounting in Economic Exchange. In: Gigerenzer, G., Selten, R. (Eds.), Bounded Rationality: The Adaptive Toolbox. MIT Press, Cambridge.

McCabe, K., Smith, V., LePore, M., 2000. Intentionality Detection and 'Mindreading': Why Does Game Form Matter? Proceedings of the National Academy of Sciences 97(8), 4404–4409.

Nelson, W.R., Jr. 2002. Equity and Intention: It's the Thought that Counts. Journal of Economic Behavior and Organization, forthcoming.

Schotter, A., Weiss, A., Zapater, I., 1996. Fairness and Survival in Ultimatum and Dictatorship Games. Journal of Economic Behavior and Organization 31(1), 37–56.

George Mason University, Interdisciplinary Center for Economic Science, 4400 University Dr.; MSN 1B2, Fairfax, VA 22030

E-mail address: {kmccabe, mrigdon1, vsmith2}@gmu.edu