Assigning Intentions when Actions are Unobservable: the Impact of Trembling in the Trust Game

James C. Cox and Cary A. Deck (corresponding author)

Running Head: Trembling in the Trust Game

James C. Cox
Department of Economics
Andrew Young School of Policy Studies
Georgia State University
P.O. Box 3992
Atlanta, GA 30302-3992
USA
(404) 651-1888
jcox@gsu.edu

Cary A. Deck
Department of Economics
University of Arkansas
402 WCOB
Fayetteville, AR 72701
USA
(479) 575-6226
cdeck@walton.uark.edu

JEL Classification: C70, C91, D64, D84

The authors gratefully acknowledge the support of the National Science Foundation. The paper has benefited from the suggestions of the anonymous referees.

This paper reports laboratory experiments investigating behavior when players may make inferences about the intentions behind others' prior actions based on higher- or lower-accuracy information about those actions. We investigate a trust game with first mover trembling, a game in which nature determines whether the first mover's decision is implemented or reversed. The results indicate that second movers give first movers the benefit of the doubt. However, first movers do not anticipate this response. Ultimately, it appears that subjects are thinking on at least three levels when making decisions: they are concerned with their own material well being, the trustworthiness of their counterpart, and how their own actions will be perceived.

1. Introduction

It is well established that individuals in laboratory games do not always pursue their own maximum monetary payoff. In the ultimatum game, materially self-interested second movers should accept any positive payoff and thus materially self-interested first movers should offer the smallest possible positive amount (or zero). However, more equal splits are frequently proposed and positive offers are often rejected.[1] In an attempt to understand such behavior researchers hypothesize that subjects may attribute intentions to actions.[2] For example, a minimal proposal might be considered greedy by a second mover, thus prompting rejection as a form of punishment. A first mover may avoid making a minimal proposal in order to try to avoid rejection, or because of altruistic preferences, or both.

Berg, Dickhaut, and McCabe (1995) introduced the investment game, which unlike the ultimatum game has a mutually-beneficial, cooperative outcome. In the investment game and the related trust game described below, a first mover can forgo a certain payoff in favor of a larger total payoff that will be allocated by a paired second mover. While material self-interest predicts that the second mover will keep everything, and thus the first mover should not forgo

the certain payoff, more cooperative behavior is frequently observed. Berg, *et al*. found that 28 of 32 first movers sent more than the minimum positive amount of money and 11 of the 28 paired second movers returned a greater amount than was received. This led Berg, *et al*. to conclude that cooperation was a "primitive" aspect of behavior. Ortmann, Fitzgerald, and Boeing (2000) among many others find the behavioral pattern to be robust to various treatments. However, Cox and Deck (2005) find that the level of social distance in the experiment protocol can affect second mover behavior in the trust game.[3] Similarly, in ultimatum games negative reciprocity or punishment is found to be dependent upon the context in which the game is played.[4]

This literature suggests that choices made by individuals depend in part on the perceived intentions of other players and how people expect their decisions to be interpreted.[5] Generous actions by the first mover in the investment game are often interpreted as trusting while the second movers' sharing the larger total payoff is often interpreted as positive reciprocity. However, to directly examine the significance of motives one needs to verify that behavior differs for the same nominal payoff decision in different circumstances. To accomplish this one can decompose a game into a series of related games, some with and some without the hypothesized motivation (see Cox 2004).[6] For example Cox and Deck (2005, 2006) examine the trust game and the associated dictator game in which the decision maker faces the same nominal choices as the second mover in the trust game, as shown in Figure 1.

In Figure 1, the numbers at the end of a branch are the dollar payoffs. In the trust game tree, the top (bottom) number is the first (second) mover's payoff. In the dictator game tree, the bottom (top) is the dictator's (other subject's) payoff. The number at the middle of each branch is the number of subjects who made that choice. Note that 8 of 24 ($\cong 33\%$) of the dictators

chose "cooperate" in the dictator game and 21 of 33 ($\cong 64\%$) of the second movers chose "cooperate" in the trust game. *Together*, data from the trust and dictator games support the conclusion that there is significant play motivated by positive reciprocity in the trust game. McCabe, Rigdon, and Smith (2003) find essentially the same pattern in a similar design in which first movers were required to play down ("choose" trust) in the involuntary trust game (see Figure 2).

It is clear from previous research that perceived intentions can affect behavior.[7] Data reported by Bohnet, Frey and Huck (2001) in a variation of the trust game framed as a contract demonstrates a similar pattern. In their game the decision to defect by a second mover, *i.e*. not perform according to the contract, is followed by a node at which with some probability nature, *i.e*. the "legal system," would detect defection and enforce the cooperative outcome minus a penalty for the second mover. While the findings are based upon repeated play environments, the general result is that for both small and large probabilities of a second mover being penalized for defection there is considerable cooperation (75% and 86% respectively), but this is not true for the intermediate case (30% cooperation).[8] In the low detection case one would expect at least as much cooperation as in the trust game given the repeated nature of the experiment and in fact it was. In the high probability case, everyone regardless of their level of self-interest should cooperate and most did. However, the intermediate case is basically a dictator control treatment in that the first mover's decision is not a signal of trust because the legal system is sufficiently developed that a risk-neutral, self-interested first mover would not choose to exit. Thus one would expect behavior similar to a dictator game which it was.

In all of the experiments discussed above subjects had complete information about the payoffs of all players and there was no uncertainty about what actions were selected.[9] However,

in many naturally occurring situations players do not have such complete information. At one extreme where players know only their own payoffs from each possible outcome, McCabe, Rassenti and Smith (1998) find that behavior closely matches the subgame perfect Nash equilibrium prediction for fully rational, exclusively self-interested agents. Of course, in that setting it is not clear how to interpret the motivations or intentions of the other players.

Our paper takes a complementary approach in which subjects have complete information about payoffs but there is uncertainty about what actions have been selected by others. That is, we ask whether people are willing to give others the benefit of the doubt. To explore this issue we consider a version of the trust game, referred to as the trembling game, where there is an exogenously determined chance that the first mover's decision is reversed. When a second mover is called upon to make a decision in the trembling game there is some probability that the first mover did not actually trust the second mover. It is our hypothesis that people who would select cooperate in the dictator game control treatment would do so in the trembling game as well.[10] Similarly, we hypothesize that those who would select defect in the trust game would also select defect in the trembling game.[11] Hence we expect this type of uncertainty to affect only those motivated by reciprocity, and therefore the percentage of people cooperating in the trembling game should be at least as great as in the dictator game but no greater than in the trust game. This approach allows us to identify how unambiguous the connection between the decision task and the perceived intention behind another's action leading to that task has to be in order to induce reciprocity.

2. Experimental Design

One hundred and twenty subjects played the trust game shown in Figure 1, with a 25% chance that the first mover's action would be reversed. Subjects were recruited from undergraduate classes and were paid a $5 show up fee in addition to the payment determined by the outcome of the one shot game. Each subject participated in only one session and had not previously participated in any similar "fairness" experiments in our laboratory. The experiment procedures followed identically those of Cox and Deck (2005) except that a sheet of additional directions was distributed to each participant.[12] This sheet read as follows.

Additional Directions:

Once a decision-maker 1 has made a decision by clicking on a branch and pressing send, that decision-maker 1 will be prompted by the computer to pick a number between 1 and 4 including 1 and 4. After all decision-maker 1s have selected a number, the experimenter will randomly draw a ball from a bingo cage. If the number the experimenter draws ***does not match*** *the number decision-maker 1 selected, then decision-maker 1's* ***decision will remain unchanged****. However, if the number drawn by the experimenter* <u>*is the same*</u> *as the number selected by decision-maker 1, then decision-maker 1's* <u>*choice will be reversed*</u> *by the computer. Decision-maker 2 will never know the number selected by the decision-maker 1 counterpart.*

After these additional directions were read aloud, subjects were able to ask questions about this procedure. Also, the bingo cage and the numbered balls were shown to the subjects, and there was a trial drawing with explanation. Thus, a second mover knew that if her paired first mover chose "trust" there was a 75% chance it would not be reversed and if her counterpart chose "exit" there was a 25% chance that it would be reversed.

If second movers give first movers the benefit of the doubt, behavior should be the same as in the regular trust game (64% cooperation). If, on the other hand, imputation of intentions has to be certain to generate reciprocal behavior, one would find less cooperation than in the regular trust game. If reciprocity is fragile then behavior in the (trust game with) trembling treatment would be similar to behavior in the dictator control treatment (33% cooperation). However, an obvious difference between the trembling treatment and the dictator treatment is the possibility of both players receiving $10. Based on previous work, such a difference could affect second mover behavior.[13] Therefore, we conducted another control treatment in which the move at the first node was determined randomly, by the flip of a coin. Following the same procedures, sixty-four additional inexperienced subjects were recruited. For this treatment the sheet of additional directions read as follows.

> *Additional Directions:*
>
> *A Decision Maker 1 has no decision to make. The branch selected at the Decision Maker 1 node will be determined by a coin flip. Decision Maker 1's will leave this room, before the coin is flipped, and return to the sign-in room.*

After these additional directions were read aloud, subjects were able to ask questions about this procedure and were given the opportunity to inspect the coin. At this point the first movers returned to the show up room and then everyone was shown the decision tree for the trust game in Figure 1 and told which branch would be selected by heads or tails.[14]

3. Results

The data from the experiments are presented in Table 1 along with the data from Cox and Deck (2005, 2006) for comparison. Two findings are readily apparent from the table. First, behavior of second movers is virtually identical in the coin flip and dictator treatments. One cannot reject the null hypothesis that the two proportions are the same given the p-value of 0.908 for the two-sided z-test.[15] Given that the two treatments generate the same behavior, subsequent analysis combines the data from these two treatments. This finding is yet more evidence of the robustness of behavior in dictator games. Recall that 33% to 35% of subjects being cooperative was also reported by Bohnet, Frey and Huck (2001) and McCabe, Rigdon, and Smith (2003) for their similar treatments discussed above.

The second finding is that as hypothesized the cooperation rate in the trembling treatment nominally falls between the cooperation rates in the regular trust game and the dictator/coin flip treatments. Of the 20 second movers in the trembling game who had an opportunity to choose, 11 (or 55%) chose cooperate. This is significantly more cooperation than the 15 of 44 (or 34%) who cooperated in the dictator/coin flip games (p-value of 0.057 in the one-tailed z-test of equal proportions against the one-sided alternative).[16] In contrast, the difference in cooperation between the trust and trembling games is not significant (p-value of 0.267 in the one-tailed z-test of equal proportions against the one-sided alternative).[17] This suggests that while there is some nominal slippage in cooperation, reciprocal behavior is not fragile in this context. That is, people who are conditional cooperators are willing to give others the benefit of the doubt by acting as though others have behaved in a manner warranting cooperation even if there is no certain evidence of that behavior.

A third finding is that even though second movers do not significantly change their behavior with the introduction of trembling, first movers do. In the trembling game only 20 of

59 (or 34% of) first movers selected trust.[18] This is significantly different from the 33 of 66 (or 50%) selecting trust in the regular version of the trust game (with a p-value of 0.069 in a two-tailed test).[19] That first movers are divided approximately equally between the two options is a fairly robust finding. For example, McCabe, Rigdon and Smith (2003) report that 17 of 27 first movers trusted, behavior not significantly different from 50%. Previous studies have not found a treatment effect for first movers. For example, Cox and Deck (2005) find that switching from a low social distance protocol where subjects are paid in person to a high social distance protocol where payoffs are double blind leads to a significant decrease in cooperation but no corresponding decrease in trust.[20]

The implication of behavior in the trembling game is that first movers seem to expect second movers to respond to the uncertainty by not giving the first movers the benefit of the doubt. Consider the decision faced by first mover *j* in the trust game. Suppose she believes that the second mover will select cooperate with probability p_j. If she picks exit her utility will be $u_j(10,10)$ while the choice of trust will result in an expected utility of $p_j u_j(15,25) + (1-p_j)u_j(0,40)$. Normalizing so that $u_j(0,40) = 0$, individual *j* will choose trust in the trust game if $p_j > u_j(10,10)/u_j(15,25)$. For a materially self-interested, risk-neutral person this simplifies to $p_j > 10/15 = 66\%$, which is close to the 63.6% percent cooperation rate observed in the trust game. Now consider the decision faced by first mover *j* in the trembling game. Suppose that she believes the second mover will select cooperate with probability π_j if given the opportunity. Then if she picks exit her expected utility will be $0.75u_j(10,10) + 0.25[\pi_j u_j(15,25) + (1-\pi_j)u_j(0,40)]$ while the choice of trust will result in an expected utility of $0.75[\pi_j u_j(15,25) + (1-\pi_j)u_j(0,40)] + 0.25u_j(10,10)$. Again normalizing so that

$u_j(0,40) = 0$, we find that individual *j* will choose trust in the trembling game if $\pi_j > u_j(10,10)/u_j(15,25)$. Thus first movers are predicted to be more or less trusting in the trust game than in the trembling game depending only on their expectations (p_j or π_j) for second mover behavior in the two games.

While the observed cooperation rate in the trembling game of 55% is not significantly lower than the cooperation rate of 63.6% in the trust game, it is nominally lower. The significantly lower occurrence of the trust branch choice in the trembling game may be a reflection of the lower nominal rate of cooperation in that game, but it may instead be an addition to the growing literature on backwards induction failure in which first movers fail to correctly anticipate the subsequent behavior of second movers (Camerer, 2003).

4. Conclusion

Often, when people make decisions they have imperfect information about the intentions behind another's perceived action and therefore have to make inferences from what they observe. To explore the ramifications of this type of uncertainty in a controlled laboratory setting we introduce trembling into the trust game. In this variant of the game, there is a 25% chance that the first mover's action will be reversed. This design allows us to address two important aspects of economic behavior. How do people behave when they have to consider how the results of their imperfectly-observable actions will be perceived rather than how their actual actions would be perceived if they were perfectly observable? How do people react when an observed action may or may not have been intentional?

In response to the second question, we find that second movers are willing to give the first mover's the benefit of the doubt. Behavior for second movers is approximately the same in

the trembling game as in the standard trust game, with a majority of the second movers who are given the opportunity to make a decision deciding to cooperate. Further, it is clear that second movers are considering the intentions of the first movers. When the first mover decision is determined randomly by the flip of a coin, the cooperation rate falls to one-third. This is approximately the same cooperation rate observed in the dictator control treatment for the trust game as well as similar treatments reported by Bohnet, Frey, and Huck (2001) and McCabe, Rigdon, and Smith (2003).

Even though second movers are willing to give the benefit of the doubt, the answer to the first question is that first movers do not anticipate this, as evidenced by the significant decrease in trust resulting from the introduction of trembling. This does not necessarily mean that first movers expect a large shift in second mover behavior. In the trust game a risk neutral materially self-interested first mover is basically indifferent between trusting and exiting given the observed behavior of second movers. Thus any reduction in anticipated cooperation could lead these first movers to strictly prefer not trusting. Of course subjects with altruistic preferences or ones that misperceives the risk of defection may continue to choose trust in the trembling treatment. Ultimately, the impact of first movers anticipating less cooperation is that substantially fewer pairs reach the cooperative outcome.

References

Berg, Joyce, John W. Dickhaut, and Kevin A. McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 122-42.

Bohnet, Iris, Bruno S. Frey, and Steffan Huck. 2001. More order with less law: on contract enforcement, trust, and crowding. *American Political Science Review* 95: 131-44.

Bolton, Gary E., and Axel Ockenfels. 2000. ERC: a theory of equity, reciprocity and competition. *American Economic Review* 90: 166-93.

Camerer, Colin, 2003. *Behavioral game theory: experiments in strategic interaction.* Princeton, N.J.: Princeton University Press.

Cox, James C. 2002. Trust, reciprocity, and other-regarding preferences: groups vs. individuals and males vs. females. In *Advances in experimental business research, edited by* Rami Zwick and Amnon Rapoport. Boston, MA: Kluwer Academic Publishers.

Cox, James C., 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46: 260-81.

Cox, James C., and Cary A. Deck. 2005. On the nature of reciprocal motives. *Economic Inquiry* 43: 623-35.

Cox, James C., and Cary A. Deck. 2006. When are women more generous than men?" University of Arkansas working paper.

Cox, James C., and Vjollca Sadiraj. 2005. Direct tests of models of social preferences and a new model. University of Arizona working paper.

Deck, Cary A. 2001. A test of behavioral and game theoretic models of play in exchange and insurance environments. *American Economic Review* 91: 1546-55.

Engelmann, Dirk, and Martin Strobel. 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review* 94: 857-69.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2003. On the nature of fair behavior. *Economic Inquiry* 41: 20-6.

Fehr, Ernst, and Klaus M. Schmidt. 1999. A theory of fairness, competition and Cooperation. *Quarterly Journal of Economics* 114: 817-68.

Güth, Werner, Steffen Huck, and Wieland Müller. 2001. The relevance of equal splits in ultimatum games. *Games and Economic Behavior* 37: 161-9.

Güth, Werner, and Reinhard Tietz. 1990. Ultimatum bargaining behavior: a survey and comparison of experimental results. *Journal of Economic Psychology* 113: 417-49.

Hoffman, Elizabeth, Kevin A. McCabe, Keith Shachat, and Vernon L. Smith. 1994. Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior* 7: 346-80.

McCabe, Kevin A., Stephen J. Rassenti, and Vernon L. Smith. 1998. Reciprocity, trust, and payoff privacy in extensive form bargaining. *Games and Economic Behavior* 24: 10-24.

McCabe, Kevin A., Mary Rigdon, and Vernon L. Smith. 2003. Positive reciprocity and intentions in trust games *Journal of Economic Behavior and Organization* 52: 267-75.

McCabe, Kevin A., and Vernon L. Smith. 2000. A comparison of naïve and sophisticated subject behavior with game theoretic predictions. P*roceedings of the National Academy of Sciences* 97: 3777-81.

McCabe, Kevin, Vernon Smith, and Michael LePore. 2000. Intentionality detection and 'mindreading': why does game form matter? *Proceedings of the National Academy of Sciences* 97: 4404-09.

Ortmann, Andreas, John Fitzgerald, and Carl Boeing. 2000. Trust, reciprocity, and social history: a re-examination. *Experimental Economics* 3: 81-100.

[1] See Güth and Tietz (1990) for a survey of experimental results in ultimatum games.

[2] An alternative approach is to model utility as depending upon the monetary outcomes of multiple players; see for example Bolton and Ockenfels (2000) and Fehr and Schmidt (1999). However, these models have had only limited success; see Deck (2001), Engelmann and Strobel (2004), and Cox and Sadiraj (2005).

[3] Specifically, Cox and Deck (2005) find that second mover behavior is less cooperative when the experimenter does not know the identity of the subjects than when subjects receive their payoffs in person from the experimenters.

[4] Responses to greedy proposals in ultimatum and mini-ultimatum games have been found to depend on the decision context and payoff structure (see Hoffman, McCabe, Shachat and Smith 1994 and Cox and Deck 2005).

[5] See McCabe, Smith and LePore (2000) for a general discussion of intentions and "mind reading."

[6] Alternately, one can make comparisons by presenting a subject with a collection of similar games and observing how choices vary with the games' payoff structures (see Falk, Fehr and Fischbacher, 2003; Güth, Huck and Müller, 2000).

[7] As noted by McCabe, Rigdon and Smith, this demonstrates a critical shortcoming in purely outcome-based models such as offered by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999).

[8] This discussion is based upon data from the last period in which different treatments involved different probabilities of detection and omits the session with fixed subject matching as those subjects knew they would continue to interact with each other in subsequent rounds.

[9] Bohnet, Frey and Huck (2001) create uncertainty about the outcome for a given action by introducing a lottery, but there is no uncertainty about what action was undertaken.

[10] As discussed in the next section, there is an alternative specification for a game that one could argue is a more appropriate baseline than the dictator game. As revealed in section 3, behavior in the dictator game and the alternative specification is indistinguishable.

[11] A referee suggested an alternative hypothesis: a second mover may feel bad that a first mover's decision could be reversed and this could make the second mover more likely to cooperate.

[12] A copy of all directions and handouts used in this study are available from the authors upon request.

[13] For example, Falk, Fehr and Fischbacher (2003) find that payoffs along other branches of the game tree can impact behavior.

[14] In half the sessions heads corresponded to the exit branch and in the other half heads corresponded to the trust branch. The terms trust, exit, cooperate, and defect were never used in interactions with the subjects, instead neutral language was used throughout. Participants were referred to as decision makers and not players or movers.

[15] In the case of comparing two sample proportions, a z-test gives the same p-value as a ^{2} test. Here we use a two-sided test because we have no prior belief about how behavior may differ between these two treatments

[16] A one-tailed test is used for second mover behavior because previous research suggests a direction for the treatment effect as discussed in the introduction. The treatment effect would not be significant based upon a two-tailed test.

[17] The treatment effect would not be significant based on a two-tailed test either.

[18] The response from one subject in the role of a first mover was omitted from the analysis because that subject had previously participated in a trust game experiment. Including this observation would not change the substantive conclusions drawn in the paper.

[19] In this case previous work does not suggest a specific direction for the alternative hypothesis and thus the question of interest is simply if there is a difference.

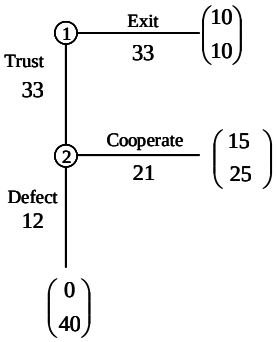
[20] Ortmann, Fitzgerald and Boeing (2000) find a lack of treatment effect on first mover behavior in a study of the investment game.

Table 1. Observed Behavior in the Four Games

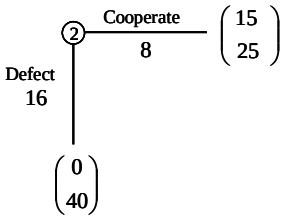
Game	Trust	Trembling	Coin Flip	Dictator
Number Selecting Cooperate	21	11	7	8
Number Selecting Defect	12	9	13	16
Cooperation Percentage	63.6%	55%	35%	33.3%

Figure 1. Frequency of Actions Chosen in Cox and Deck (2005, 2006)

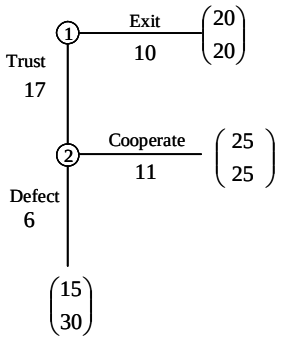
Figure 2. Frequency of Actions Chosen in McCabe, Rigdon and Smith (2003)



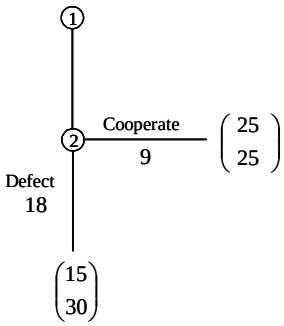
Trust Game



Dictator Game



Voluntary Trust Game



Involuntary Trust Game