

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

NOTE TO USERS

This reproduction is the best copy available.

UMI

**COOPERATION AND INTENTIONS IN EXPERIMENTAL
BARGAINING GAMES**

by
Mary L. Rigdon

Copyright © Mary L. Rigdon 2001

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF ECONOMICS
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY
In the Graduate College
THE UNIVERSITY OF ARIZONA

2001

UMI Number: 3023524

**Copyright 2001 by
Rigdon, Mary Lisa**

All rights reserved.

UMI[®]

UMI Microform 3023524

Copyright 2001 by Bell & Howell Information and Learning Company.

**All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.**

**Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346**

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Marji

ACKNOWLEDGMENTS

In reflecting on all the people I would like to thank, I realize just how fortunate I am. I begin by thanking Bob Basman and Stan Masters, whose insight and mentorship while I was at SUNY-Binghamton shaped the way I approach economics. Stan's influence has kept me interested in Labor Economics. Bob is responsible for first kindling my interest in Experimental Economics, and encouraged me to leave New York for the experimental hotbed that is Arizona. Thanks to James Marchand, Richard Chapman, the late Ann Hansen, and Susan Cottler: all provided me an invaluable learning experience during my undergraduate training. The Economic Science Laboratory provided me the financial and moral support that I needed in carrying out this research. Thanks to Rachel Croson, Catherine Eckel, Eline Van der Heijden, Martin Dufwenberg, Urs Fischbacher, Dan Houser, Massimo Piatelli-Palmerini, participants at the Economic Science Association meetings in Tucson (October 2000) and in Amsterdam (October 2000) for comments on earlier versions of ideas contained in this dissertation. The National Science Foundation and the Russell Sage Foundation provided funding for a majority of the experiments contained in these chapters. Thanks also to my parents who showed me that it is possible to live through a dissertation (or even two). Thanks to my beagle boys, Jasper and Attila, who kept me company in my office as this dissertation took shape.

Thanks to my committee members: Vernon Smith, Kevin McCabe, and Dan Houser. They made this process more bearable. Vernon and Kevin have served as friends, mentors, and advisors during my time at the University of Arizona. They are the best example that economics can be a real scientific endeavor. Without this example, and their unflinching support of my work, my education would have been incomplete. Chapters 2 and 3 represent collaborations with them.

And, most importantly, thanks to Thony. His support, encouragement, and sense of humor through all (well, and \LaTeX typesetting help too) has meant the most. This dissertation is dedicated to him.

TABLE OF CONTENTS

LIST OF FIGURES	6
LIST OF TABLES	7
ABSTRACT	8
CHAPTER 1. PERSONAL EXCHANGE AND BARGAINING EXPERIMENTS . .	10
1.1. The Behavioral Bargaining Problem	10
1.2. Background	11
1.3. Early Experiments	13
1.4. Trust Games	14
1.5. Explaining Bargaining Behavior	16
CHAPTER 2. POSITIVE RECIPROCITY AND INTENTIONS IN TRUST GAMES	19
2.1. Introduction	19
2.2. Trust and Reciprocity	21
2.3. Outcome-based Models	23
2.4. Localized Play in an Extended Trust Game	24
2.5. Experimental Design and Procedures	26
2.6. Predictions and Hypotheses	27
2.7. Results	29
2.8. Discussion	31
CHAPTER 3. SUSTAINING COOPERATION IN TRUST GAMES	34
3.1. Introduction	34
3.2. Sustaining Trust	35
3.3. Experimental Design and Procedures	39
3.4. Results	40
3.5. Conclusion	48
APPENDIX A. INSTRUCTIONS FOR VOLUNTARY/INVOLUNTARY EXPERIMENTS	49
APPENDIX B. INSTRUCTIONS FOR SORTING EXPERIMENTS	53
REFERENCES	58

LIST OF FIGURES

FIGURE 1.1. Two-person trust game	15
FIGURE 1.2. Trust game: Single-Play by Undergraduates	15
FIGURE 2.1. Single Play Results (McCabe <i>et al.</i> , 1998)	25
FIGURE 2.2. Voluntary trust game	26
FIGURE 2.3. Involuntary trust game	27
FIGURE 2.4. Frequency of moves in the Voluntary trust game	30
FIGURE 2.5. Frequency of moves in the Involuntary trust game	30
FIGURE 2.6. Percent of Player 2s Cooperating in One-Shot Game	31
FIGURE 3.1. Trust game	35
FIGURE 3.2. Percent of Players 1 Choosing SPE Over Time	42
FIGURE 3.3. Percent of Players 1 Trusting Over Time	43
FIGURE 3.4. Percent of Players 2 Cooperating Over Time	43
FIGURE 3.5. Percent of Players 2 Defecting Over Time	44
FIGURE 3.6. Average Efficiency Scores Over Time	46
FIGURE 3.7. Cooperators versus Non-cooperators: Percent of Each Type Reaching the Cooperative Outcome of (25, 25)	47

LIST OF TABLES

TABLE 3.1.	Experimental Design	40
TABLE 3.2.	Conditional Outcomes by Trial Block	41
TABLE 3.3.	Trust Score by Condition	44
TABLE 3.4.	Efficiency Score, ν , by Condition	46
TABLE 3.5.	ANOVA for Efficiency Score ν	46

ABSTRACT

The Behavioral Bargaining Problem poses a trio of questions: (1) How do real economic agents behave in bargaining environments?; (2) Why do they behave the way they do?; and (3) What conditions/institutions sustain that behavior? This dissertation is about experimental results which suggest answers to each of these questions. Chapter 1 is a brief overview of early experiments which address how economic agents behave in bargaining environments. Under a wide variety of conditions a significant proportion of subjects' behavior deviates from that predicted by non-cooperative game theory. Chapter 2 tests several theories from behavioral game theory which aim at explaining why subjects cooperate in 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. I report laboratory data from simple two-person trust games which is inconsistent with outcome-based models, but predicted by the trust and reciprocity hypothesis. Chapter 3 is devoted to exploring one way of sustaining cooperative behavior in a simple two-person trust environments. It is well-known in evolutionary game theory that "clustering" in a population can allow an evolutionary stable strategy to be invaded in a finitely repeated Prisoners' Dilemma game. This idea can be adapted to bargaining by noting that an agent's history of choices gives him a track record. Players can be typed based on their recent track record as whether or not they are trusting (for Players 1), and whether or not they are trustworthy (for Players 2). Once the players are typed, they can then be paired according to those types: trustors with trustworthy types, and similarly non-trustors with untrustworthy types. This sort of matching protocol induces clustering within the population. The empirical ques-

tion is whether this adaptation of clustering to bargaining environments can sustain cooperative play analogous to the situation in finitely repeated Prisoners' Dilemma games. The results indicate that such a sorting mechanism allows cooperative play to emerge and, once it emerges, sustains it over time.

Chapter 1

PERSONAL EXCHANGE AND BARGAINING EXPERIMENTS

1.1 The Behavioral Bargaining Problem

Jane is looking to buy herself a bicycle, and she is willing to pay a maximum of \$100 for a used bike. John, as it happens, is moving to a distant city and wants to sell his bike. But John is not much of a salesman: if he does not sell his bike to Jane he will just abandon it. It is common knowledge between the two that Jane is willing to pay \$100 for the bike, and that if John does not sell to Jane the bike is worthless to him. Time is short for both of them—John is leaving town, and Jane needs a bike soon. In fact, the constraints only allow Jane to make an offer and John to either accept or decline without making a counteroffer. Jane and John must *bargain* over the potential sale of the bicycle.

John and Jane are involved in a *bilateral bargaining situation*. A bilateral bargaining situation is any situation where two parties have the opportunity to interact in order to reach a mutual gain in more than one way. In bargaining situations like the one above the parties attempt to distribute the potential mutual gain between them. Firm–employee relationships, and buyer–seller relationships on the Internet are also concrete examples of bilateral bargaining environments. It is easy to identify bargaining situations, and almost trivial to say that understanding them is important to understanding economic behavior. But there are two associated problems with bargaining environments. The first is a theoretical/normative problem: How would ideally rational agents interact in a given bargaining situation? This is *The Bargaining Problem*, and the canonical solution that Nash (1950) provided has paved the way for axiomatic bargaining theory.

But there is a second bargaining problem in the vicinity—the *Behavioral Bar-*

gaining Problem. This problem, which is an empirical one, poses a trio of questions: (1) How do real economic agents behave in bargaining environments?; (2) Why do they behave the way they do?; and (3) What conditions/institutions sustain that behavior? This is the bargaining problem that will take center stage here.

The rest of this chapter will address the first question of the Behavioral Bargaining Problem. Anticipating just a bit, the first major result from research in behavioral bargaining is that the solution to the theoretical problem does not adequately predict behavior in bargaining experiments. The trend is rather general: the theoretical solution to bargaining problems comes in the way of solution concepts of non-cooperative game theory, but behavior in a long series of experiments (since 1982) indicates that significant proportions of subjects deviate from this prediction—and deviate for mutual gain. The chapter closes with some hints at an answer to question (2). Subjects. I contend, read and send signals about their intentions to cooperate in bargaining environments. This sort of explanation is tested, in Chapter 2, against other recent models of bargaining behavior. Chapter 3 reports the results of experiments which aim at answering the third question of the Behavioral Bargaining Problem. Cooperative behavior can be sustained by sorting subjects by their history of cooperative play, and matching them for future interactions accordingly. Appendix A and Appendix B contain the subject instructions from the experiments reported in Chapters 2 and 3, respectively.

1.2 Background

Since Nash's seminal work, it has become common in economics to model bilateral bargaining environments as a special case of two-person noncooperative games. For completeness, I provide here a brief overview of some of the tools and concepts of game theory that are most centrally at issue in the Behavioral Bargaining Problem.¹

¹This background information is meant only as a primer on the classical treatment of extensive form games. See Fudenberg and Tirole (1991) for a thorough introduction.

An *extensive form game of perfect information* is a graph-theoretic tree—i.e., a collection of nodes (the *decision nodes* and the *outcomes*) and directed, acyclic edges between the nodes (*branches*). We will only be concerned with two-person games. The non-terminal nodes are labeled with player numbers; node i is Player i 's decision node, a node at which she must make a decision. Terminal nodes are marked with *outcomes*—an ordered tuple representing what each player would get if that outcome were to be reached. The branches are labeled with *actions*. If node i has two descending branches, say one labeled A and one labeled B , then the interpretation is that Player i has a choice at that node between A and B — A and B are actions that are *open* to i at that node. The games are perfect information games in that the history of moves is transparent between the players, and (in principle) no one ever forgets what has happened. I am most interested in pure bargaining environments which can be represented by extensive form games of perfect information. So, without risk of ambiguity, we may simply refer to these structures as *games*.

The central solution concept for extensive form games is the concept of *subgame perfection*. A *strategy profile* s is a function on a game which assigns to every decision node an action (from among the actions that are open at that node). Each profile determines a unique *outcome*.² Let $O(s)$ be the outcome that s determines. A profile s^* is a Nash Equilibrium of a given game iff $O(s_i^*, s_i^*) \succeq_i O(s_i^*, s_i)$ for every s_i of Player i ($i \in \{1, 2\}$). A profile s^* is a *Subgame Perfect Equilibrium* in a game G iff for every subgame G' of G s^* (restricted to the nodes in G') is a Nash Equilibrium of G' . An outcome of an extensive form game is subgame perfect iff it is the outcome determined by a subgame perfect strategy profile.

Every two-person extensive form game of perfect information has a subgame perfect outcome. If the players' preferences are strict, then it is a unique subgame perfect outcome. Moreover, all of the bargaining situations we will encounter can be represented as extensive form games of this sort. Hence, all of the bargaining games we

²Since games are acyclic, a path through a game tree is unique.

will see have unique subgame perfect outcomes. From the point of view of the Behavioral Bargaining Problem this is significant: noncooperative game theory makes a clear prediction about behavior in these environments. As we will see, the prediction is not adequately supported by data from bargaining experiments.

1.3 Early Experiments

A particularly simple bargaining environment is the ultimatum game. In a two-person ultimatum game, a proposer (Player 1) offers a distribution of c dollars to a responder (Player 2) and $M - c$ dollars for himself. Following the offer, Player 2 must either accept (in which case the distribution stands) or reject (both players receive zero dollars). If the responder is rational then she will accept any offer where $c > 0$, and so a rational proposer should offer the smallest such c . Such a state of affairs in a given ultimatum game is the unique Nash equilibrium of that game.

Güth, *et al.* (1982) was the first experimental test of the ultimatum game. Subjects in their study were randomly assigned the role of proposer or responder. All subjects could see each other, but none knew which of the others was his counterpart. Offers were significantly larger than the smallest unit of account, $c = \$1$. Furthermore, responders did not hesitate to reject small offers. Offers were on average more “fair”, and acceptance rates of “fair” offers exceeded those for “unfair” offers.

Forsythe, Horowitz, Savin, and Sefton (FHSS, 1994) wanted to examine the extent to which ultimatum game results are driven by considerations of fairness, and so to test the robustness of the Güth, *et al.* results. FHSS make use of an additional game: the dictator game. Here the dictator decides how to divide the M dollars between the two players and Player 2 does not have the option to reject the distribution. As a result, a dictator game provides a control for the proposer’s expectations of rejection by the responder in an ultimatum game. FHSS argue that if fairness is the primary reason for observing non-trivial offers in the ultimatum game, then we should also

see the same offer distribution in the dictator game. They find that, contrary to the hypothesis of fairness, players are more generous in the ultimatum game than in the dictator game. This basic finding is replicated by Hoffman, McCabe, Shachat, and Smith (HMSS, 1994). For an overview of bargaining experiments, see Roth (1995).

1.4 Trust Games

This general result from dictator and ultimatum games generalizes to other, more dynamic, bargaining environments. Of particular interest is the class of two-person trust games. In a trust game, Player 1 chooses between an outside option (which is inefficient) and “investing” in a possible future mutual gain to be divided by Player 2. Player 2’s dominant action, however, leaves Player 1 with a loss compared to the value of the outside option. Player 1’s outside option is the subgame perfect equilibrium (SPE) of the game.

Consider the two-person trust game in Figure 1.1. Player 1 can choose to split \$20 evenly with his counterpart, Player 2, or choose to pass the game to Player 2. If the game is passed, Player 2 then decides between the mutually beneficial, cooperative outcome (giving \$15 to Player 1 and keeping \$25 for herself) or the selfish/defection outcome (giving Player 1 nothing and keeping the entire \$40 for herself). The SPE predicts that Player 1 will take his outside option, and therefore that the (10, 10) outcome will be reached. The argument is a familiar backward induction argument. Player 2, if she were faced with a choice between \$25 and \$40 would choose \$40 (since she prefers more to less). Player 1 knows that this is so, and so he knows that playing down would certainly yield him a payoff of \$0. He prefers \$10 to \$0, so he moves right and takes his outside option.

McCabe and Smith (2000a) report data for 24 pairs of undergraduates from the University of Arizona who participated in a one-shot anonymous interaction in a standard two-person trust game. Figure 1.2 shows the frequency of play at each

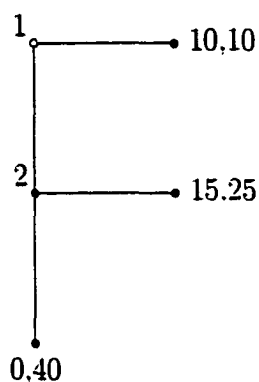


FIGURE 1.1. Two-person trust game

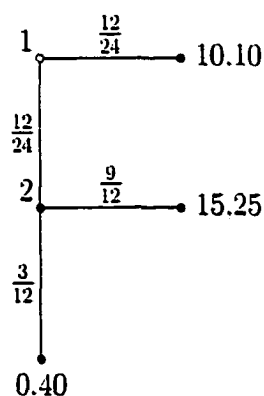


FIGURE 1.2. Trust game: Single-Play by Undergraduates

decision node. Contrary to the non-cooperative prediction that all Players 1 would move right, half of the first-mover undergraduates chose to play down. 75% of these trusting subjects experience reciprocity by their counterpart. The realized expected payoff to a Player 1 for choosing down was $0.75 \times 15 = \$11.25$. This is a gain over the SPE, \$10. It is worth noting that despite the half who deviate from the non-cooperative prediction, the game theoretic prediction does explain half of the data.

Berg, Dickhaut, and McCabe (BDMc, 1995) use a similar trust game that has a much richer message space. In particular, in this investment game, all subjects begin

with a \$10 show-up fee. They are divided into two rooms, A and B. A subject in room A (the Sender) decides how much of his show-up fee—\$0, \$1, \$2, . . . \$10—he would like to send to his anonymous counterpart in room B. It is publicly announced that any amount sent on is tripled by the experimenter. A room B subject (the Receiver) then decides how much of this tripled amount she would like to keep and how much she would like to send back to her counterpart in room A.³ If players are only concerned about their own material well-being, then any amount subjects in room B receive, they will keep. So subjects in room A, knowing this, will decide via backward induction to send zero dollars. This is the unique Nash equilibrium of the investment game. However, subjects' behavior does not match this prediction. In fact, "30 of 32 room A subjects sent money (\$5.16 on average)" (p. 323). Furthermore, those in room B reciprocate by sending back portions of what they receive, especially to those who are most trusting (i.e. those in room A who send all \$10 to their counterpart). For a detailed look at trust game experiments, and how theories of cooperation fare in them, see McCabe, Rigdon, and Smith (2001).

1.5 Explaining Bargaining Behavior

Early experiments suggest that a considerable proportion of play in two-person bargaining games does not arrive at subgame perfect and, more generally, Nash equilibria. Instead, a significant number of subjects arrive at cooperative outcomes. The results are, moreover, robust: these environments are thought to be particularly favorable to the game theoretic noncooperative equilibria—they are conducted as single play games so that neither a history or a future can matter within the same game. Furthermore, subjects are anonymously paired as strangers with an unknown identity. Given these data, we need a different approach to the Behavioral Bargaining Problem than the one provided by noncooperative game theory.

³It is important to note two things about this particular design: subjects were actually holding the dollar bills as they made their decisions and a double-blind protocol was used.

Game theory plays a dual role in economics. Sometimes it is presented as a normative model, what agents ought to do. But, as scientists, economists are interested in game theory as an empirically informative theory. We want two things from our theories. First, of course, they must be able to account for the data which they are meant to be about. Call this the *descriptive adequacy* of a theory. But there is a further, and more important, requirement: theories must explain why the data are the way that they are. A theory must be *explanatorily adequate*. In bargaining experiments, the noncooperative theory does not provide either a descriptively or an explanatorily adequate account of the data. A natural remedy would be to *extend* traditional game theory with solution concepts and the like to make it a more explanatorily adequate theory. This, roughly, is the project of *behavioral game theory*, and the Behavioral Bargaining Problem is one of its main targets (Camerer, 1997).

But extending game theory in this way to cope with its inadequacy in describing and predicting behavior in bargaining environments has met with resistance. Two reasons have been offered in the literature, but neither of them is compelling. The first argument runs as follows. There is no need for a new theory to explain the experimental results in dictator and ultimatum games since it is clear that people just have a taste for fairness. But, of course, by calling this behavior “fair” one has only described, and not explained, the phenomenon. Ultimately we want a theory which says why it is that subjects do what they do in bargaining environments.

The second argument—advanced by Ken Binmore (1999)—is that cooperative behavior might disappear if agents are given feedback over time. The idea is that laboratory data from inexperienced subjects are noisy, and this is because we humans are boundedly rational. By allowing subjects to play repeatedly the data in bargaining games may converge to noncooperative equilibria after all. So while we need *some* new theory, it should be a theory of learning, not a theory of personal exchange in bargaining.

As appealing as this view may be for its conservativeness, it is misguided (with

respect to exchange problems) for at least three reasons. First, as mentioned above, the single play environment gives noncooperative theory its best shot. Second, data provided by McCabe,*et al.* (1996) and McCabe,*et al.* (1998) demonstrate that cooperative behavior can be sustained under a variety of repeated trial conditions. Using a trust game where players have the opportunity to punish defection, and a matching protocol where each subject plays every other subject exactly once (reputation building cannot occur), cooperation does not deteriorate over time. There is in fact no significant tendency under repeat single play conditions for the proportion of trusting by first movers, or the use of reciprocal trustworthy responses by second movers, to unravel across time. Learning models cannot account for this data. Lastly, cooperative outcomes in repeated bargaining games are supported by the folk theorem. So while it is important to understand how boundedly rational agents adapt in economic situations, the challenges that bargaining experiments pose to standard noncooperative game theory cannot be explained away by appeal to learning in games.

The Behavioral Bargaining Problem is beginning to yield to experimental investigation. How do subjects behave in bilateral bargaining? A wide variety of experimental treatments under a variety of conditions indicate that behavior cannot be satisfactorily explained by noncooperative game theory. Why do subjects behave the way they do? The next chapter aims at discriminating between two classes of models, each attempting to explain why and how subjects cooperate in these environments. How can cooperative behavior be sustained? Chapter 3 describes a method of population clustering that, given the experimental results reported there, supports cooperative play over time in simple bilateral bargaining.

Chapter 2

POSITIVE RECIPROCITY AND INTENTIONS IN TRUST GAMES

2.1 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. But receiving the future benefit 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 that in exchange the worker will provide a greater effort (thus achieving a cooperative outcome). Such environments can be nicely modeled by *two-person trust games*.

There is ample experimental evidence which suggests a considerable proportion of play in two-person trust games deviates from that predicted by standard non-cooperative game theory (BDMc, 1995; McCabe, *et al.*, 1996, 1998). That is, 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 exclu-

sively 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, on the other hand, emphasizes the role of intentions in achieving cooperative outcomes in personal exchange. The models in McCabe and Smith (2000b), Dufwenberg and Kirchsteiger (1998), and Falk and Fischbacher (1999), 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. Since it is only the intrinsic properties of outcomes that are supposed to drive behavior, 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 2.2) gives a more detailed discussion of the trust and reciprocity hypothesis. Section 2.3 provides an overview of several recent outcome-based models. Before turning to the experimental design (Section 2.5), we first consider some of our earlier work which, while suggestive

of our view that first movers' opportunity costs matter to how second movers interpret observed actions, is not sufficient for distinguishing between intention- and outcome-based models (Section 2.4). Section 2.6 contains the predictions and hypotheses for our design and Section 2.7 reports the results.

2.2 Trust and Reciprocity

Within the class of intention-based approaches, we want to focus on the trust and reciprocity (TR) hypothesis and how it can intuitively 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, in order 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 Player 2's (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 in terms of 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. 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

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

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.

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; Carruthers and Smith, 1996; Gillies and Rigdon, 2000). 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 routinely predict how others will behave. According to 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.

2.3 Outcome-based Models

Here we briefly sketch three recent outcome-based models: ERC (Bolton and Ockenfels, 2000), the Fehr-Schmidt model (1999), and Levine's altruism/spitefulness model (1998). In Section 2.6 we will derive specific predictions for our treatments.

Bolton and Ockenfels propose in their ERC preference model for two-person games a motivation function: $v_i = v_i(y_i, \sigma_i)$ 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. The ERC model has four assumptions governing the behavior of v_i . Of particular relevance to our games below are the assumptions of narrow self-interest (NSI) and comparative effect (CE).

(NSI) Let $v_i^1(y_i^1, \sigma)$ and $v_i^2(y_i^2, \sigma)$ be two motivation functions for Player i on two outcomes where σ is a fixed relative argument. Then, if $y_i^1 > y_i^2$ then $v_i^1(y_i^1, \sigma) > v_i^2(y_i^2, \sigma)$.

(CE) Holding y_i constant, the motivation function v_i is strictly concave in σ_i , with a maximum at which i 's own share is $\frac{1}{2}$.

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 are. 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:

(FSU) $U_i(\vec{x}) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\}$

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.

Finally, Levine models the altruism and spitefulness of players by considering an adjusted utility function. The adjusted utility function (LU) for Player i in a two-person game is a function of i 's own payoff (u_i), the monetary payoff of the other player (u_j), how spiteful/altruistic i is (a_i), how spiteful/altruistic the other player is (a_j), and a fifth parameter, which acts as a scalar on the other player's altruism/spitefulness (λ).

$$\text{(LU)} \quad v_i(u_i, u_j, a_i, a_j, \lambda) = u_i + \frac{a_i + \lambda(a_j)}{1 + \lambda} u_j$$

where $0 \leq \lambda \leq 1$ and $-1 < a_i < 1$. The distributions of the a_i s and λ s are assumed to be uniform and common knowledge. Players are typed by their values of a_i and λ . As with all outcome-based models, it is the properties of outcomes—and *only* properties of outcomes—that drive explanation of behavior.

2.4 Localized Play in an Extended Trust Game

Before turning to our experimental design, in this section we briefly sketch some earlier work which serves as the empirical motivation for the voluntary and involuntary trust treatments that we report below. McCabe, *et al.* (1998) study the emergence and support of trust in extensive form bargaining environments under a variety of conditions.

Figure 2.1 reports the findings in a treatment in which subjects are randomly matched in a one-shot game. In this game there is no threat of direct punishment for

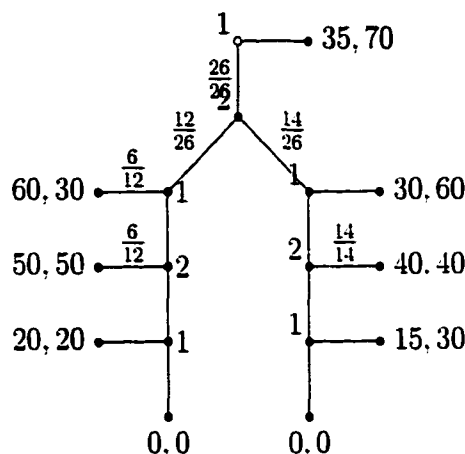


FIGURE 2.1. Single Play Results (McCabe *et al.*, 1998)

not choosing the cooperative outcome.² As shown in the figure, play centered on just three outcomes of the game: (40, 40), (60, 30), and (50, 50). When subjects played in the right subgame, the outcome in every case was the subgame perfect outcome of (40, 40). When subjects played in the left subgame, play was divided equally between the defection outcome of (60, 30) and the reciprocity outcome of (50, 50).

These results are consistent with the hypothesis that Player 1s are making an assessment of the opportunity cost between the subgame perfect outcome in the right subgame, and the trust outcome of (50, 50) in the left subgame, and a significant proportion of Player 2s interpret this as a signal to enter a reciprocal-trust relationship. However, this game alone cannot distinguish between outcome-based explanations and the TR hypothesis. The experimental treatments of the voluntary and involuntary trust games, which are highly reduced versions of Figure 2.1, allow us to directly investigate the explanatory boundary between these two approaches.

²The same paper also considers a trust/punishment game where the (60, 30) and (50, 50) are interchanged.

2.5 Experimental Design and Procedures

We now consider our two treatments below: the voluntary trust environment and the involuntary trust environment. In all experiments, subjects were paid \$5.00 for appearing on time for the experiment. At the end of the experiment, their accumulated earnings were paid to them privately (single-bind protocol). The interactions consisted of anonymous/random pairings in a one-shot computerized game. The payoffs are actual (US) dollar amounts the subjects could earn, and are common knowledge. The subjects were undergraduates at the University of Arizona and did not have prior experience with a trust environment. Each experimental session consists of 12 subjects, six in each treatment condition. As a subject enters the Economic Science Laboratory he is paid a \$5.00 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 games are played sequentially. The experiments lasted on average 30 minutes, from arrival to completion. Each subject participates in one and only one such experiment.

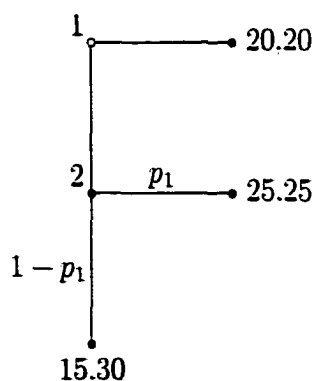


FIGURE 2.2. Voluntary trust game

The voluntary trust game is represented in Figure 2.2. Player 1 has an outside option of (20, 20) which is the subgame perfect equilibrium (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).

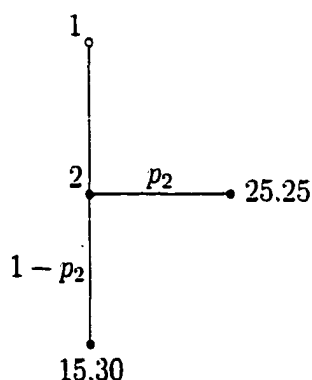


FIGURE 2.3. Involuntary trust game

Compare Figure 2.2 to the involuntary trust game in Figure 2.3. 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 voluntary trust game) and zero (in the involuntary trust game).

2.6 Predictions and Hypotheses

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

The ERC prediction is as follows. After a move down by Player 1, Player 2s 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, $p_1 = p_2$.³

For the Fehr-Schmidt prediction we need only look at the adjusted utilities of second movers. For the cooperative (25, 25) outcome, in the voluntary trust game Player 2's utility is:

$$\begin{aligned} U_2^{vtg}(25, 25) &= 25 - \alpha_2 \max\{25 - 25, 0\} - \beta_2 \max\{25 - 25, 0\} \\ &= 25 \end{aligned}$$

For the involuntary trust game, Player 2 has exactly the same choices and possible outcomes, and so the same adjusted utility of the (25, 25) outcome:

$$\begin{aligned} U_2^{itg}(25, 25) &= 25 - \alpha_2 \max\{25 - 25, 0\} - \beta_2 \max\{25 - 25, 0\} \\ &= 25 \end{aligned}$$

The utilities for the defection outcome (15, 30) are also identical across these games:

$$\begin{aligned} U_2^{vtg}(15, 30) &= 30 - \alpha_2 \max\{15 - 30, 0\} - \beta_2 \max\{30 - 15, 0\} \\ &= 30 - \beta_2 15 \\ U_2^{itg}(15, 30) &= 30 - \alpha_2 \max\{15 - 30, 0\} - \beta_2 \max\{30 - 15, 0\} \\ &= 30 - \beta_2 15 \end{aligned}$$

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

Levine's model, in the end, has the same prediction and for the same reasons. For fixed values of a_i , a_j , and λ (which are assumed to be uniformly distributed among

³This 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.

the population), two outcomes with identical payoffs have the same adjusted utilities. Therefore, rates of cooperation by Player 2s should not differ across the involuntary and voluntary trust games (the prediction is that $p_1 = p_2$).⁴

The TR hypothesis offers a very different prediction across these treatments. In the involuntary treatment 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; i.e., such conditions significantly reduce Player 2's ability to reciprocate since she cannot reliably attribute intentions of trust to Player 1. 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 .

$$H_0 : p_1 - p_2 \leq 0$$

$$H_1 : p_1 - p_2 > 0$$

where p_1 is the proportion of moves at (25, 25) conditional on Player 1's move down in the voluntary trust game, and p_2 is the proportion of moves at (25, 25) conditional on Player 1's move down in the involuntary trust game. The predictions of outcome-based models are hence represented under our null hypothesis.

2.7 Results

Figure 2.4 records the proportion of observed play at each node in the voluntary trust treatment, and Figure 2.5 records the proportion of observed play at each node

⁴An alternative version of the Involuntary Trust Game would be to have Player 1's outside option be (0, 0). While this is certainly an interesting empirical variation, the predictions of these outcome-based models about the level of cooperative play remain unchanged.

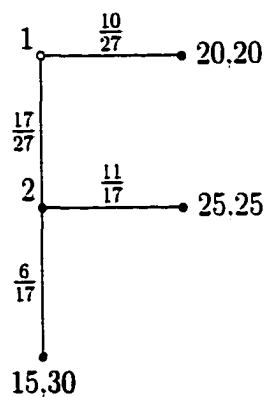


FIGURE 2.4. Frequency of moves in the Voluntary trust game

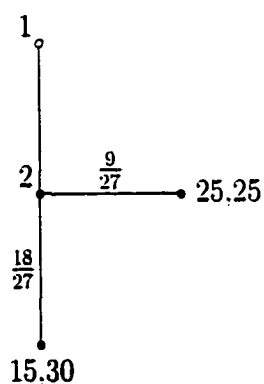


FIGURE 2.5. Frequency of moves in the Involuntary trust game

in the involuntary trust treatment. Figure 2.6 charts the comparison of the relative proportion of cooperative play.

The null hypothesis that the proportion of cooperative outcomes across treatments remains constant is easily rejected by both a normalized z -score and bootstrap test (p -value < 0.01). There is, therefore, a significant treatment effect between the two environments.

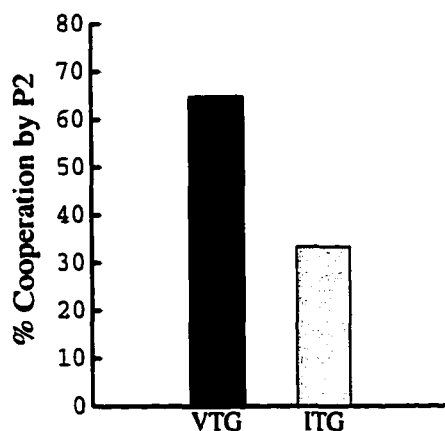


FIGURE 2.6. Percent of Player 2s Cooperating in One-Shot Game

2.8 Discussion

The data in these simple experiments are inconsistent with the predictions of the ERC, Fehr-Schmidt and Levine 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 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 hypothesis that cooperative play occur 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 since Player 1 had no choice but to move down the tree.

The data reported here are not the only 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 (HMSS, 1994; HMS, 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. Similarly, choice behavior also differs dramatically between normal and extensive form treatments (McCabe.*et al.*, 2000), but this is not consistent with outcome-based models like those considered here.⁵

A similar conclusion about the inadequacy of outcome-based models is reached by FFF (1999). 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, i.e. for the case of a left branch offer and for the case of a right branch offer, without knowing what the first mover had proposed (p.5). 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; Rapoport, 1997; Schotter.*et al.*, 1996). 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).⁶ So the experiments reported here offer a more direct test of

⁵See McCabe, Rigdon, and Smith (2001) for a further discussion.

⁶In 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 (2000). In order for such results to go through one needs the auxiliary hypothesis that responders in these games behave the same regardless of if they *actually* see the first mover's choice or are just told to *assume* that the first mover has chosen X . There is some evidence that counterfactual choice and

outcome-based models.

In the involuntary trust treatment we do observe that $\frac{9}{27}$ still end up at the cooperative outcome. What should we conclude about the motivations of these Player 2s? It is well known that the level of pure altruism is sensitive to the implementation of double blind payoffs (HMS, 1996). Since we wanted to compare the results from the voluntary/involuntary treatments with McCabe, *et al.* (1996), a double-blind protocol was not appropriate. We conjecture that the 33% of cooperative play is largely residue from our single-blind payoff procedures.

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 knowledge 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 Player 1s 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.

actual choice come apart (this is known as the Langer effect), but this is certainly an area that needs empirical testing in personal exchange environments.

Chapter 3

SUSTAINING COOPERATION IN TRUST GAMES

3.1 Introduction

There are two related problems of cooperation in bargaining environments. The first problem is to explain why people cooperate in personal exchange environments when the standard core of game theory predicts otherwise. This problem has received considerable attention in the recent literature. The second problem is to say how cooperation can be sustained once it emerges. The second problem has received comparably less attention than the first. This paper studies the effect of a treatment that controls for the history of cooperation by procedures unknown to the subjects so that cooperation is not sustained by common knowledge (expectations).

Even though sustaining cooperation has received less attention in bargaining situations, it has been a primary focus in Prisoner Dilemma (PD) and public good games (Axelrod, 1984, 1997; Ledyard, 1995). Consider the analysis of the finitely repeated PD game in Axelrod (1984). In this game, always defecting is an evolutionary stable strategy in the sense that it does not pay to cooperate in a population where everyone else always defects. Yet a small band of conditional cooperators (say, tit-for-tat players) can invade a population of unconditional defectors provided that the cooperators can *cluster*. That is, if these cooperators interact more often with each other than with the defectors, then the population can be invaded. This trades on assuming that the pairing in the interactions is not random. The problem with random pairing is that the chance of conditional cooperators meeting each other is low.

We want to adapt this idea of population clustering to a simple two-person trust game. An agent's history of choices gives him a track record. Players can be typed based on their recent track record as whether or not they are trusting (for Players

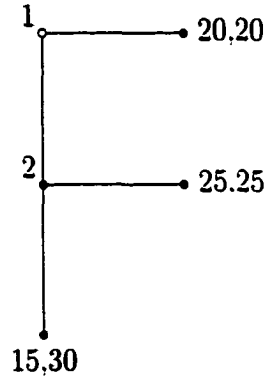


FIGURE 3.1. Trust game

1), and whether or not they are trustworthy (for Players 2). Once the players are typed, they can then be paired according to those types: trustors with trustworthy types, and similarly non-trustors with untrustworthy types. This sort of matching protocol induces clustering within the population. The empirical question that we want to address is whether this adaptation of clustering to bargaining environments can sustain cooperative play analogous to the situation in finitely repeated PD games.

In the next section we describe a two-person trust game and our mechanism for clustering the population. We then discuss the design and procedures used in our experiments (Section 3.3). Data analysis follows in Section 3.4.

3.2 Sustaining Trust

In the game pictured in Figure 3.1, Player 1 is asked to choose from the following: (1) You are given \$40, which you can split evenly with another subject in the experiment—Player 2—in which case the game is over or (2) You present Player 2 with two choices, either Player 2 can take \$30 out of \$45, leaving you \$15; or she can split \$50 evenly between the two of you.

We implemented the idea of clustering by typing players based on their observed

moves in the trust game above, and (in one treatment) match players based on their types. Types come in the form of a “trust score”, say τ where $\tau \in [0, 1]$. The idea is that each player will have a score which is a fraction with the numerator being the number of times the player cooperated and the denominator being the number of chances the player had to cooperate. At the end of each period, the algorithm begins by looping through the decisions made by all the Players 1 and calculating their respective score and then does the same for all the Players 2.

Algorithm 1 (Player 1 Trust Score). Let c_1 (d_1) indicate a cooperative (defection) move by Player 1. Then the trust score of a Player 1 after Round n , τ_n^1 , is given by the following algorithm:

1. If $n = 0$: $\tau_n^1 = 0$

2. If $n \leq 5$: Let k be the number of c_1 moves through Round $n - 1$. Then:

$$\tau_n^1 = \begin{cases} \frac{k}{n} & \text{if } d_1 \text{ in Round } n \\ \frac{k+1}{n} & \text{if } c_1 \text{ in Round } n \end{cases}$$

3. If $n > 5$: Let k be the number of c_1 moves in Rounds $n - 1, \dots, n - 4$. Then:

$$\tau_n^1 = \begin{cases} \frac{k}{5} & \text{if } d_1 \text{ in Round } n \\ \frac{k+1}{5} & \text{if } c_1 \text{ in Round } n \end{cases}$$

That the divisor (when $n > 5$) is always 5 puts a premium on the last five interactions of the players. Pre-theoretically, there is a recency effect of goodwill—recent acts of goodwill overshadow distant acts of ill-will. The trust score algorithm for Player 1 codifies this intuition by only keeping track of the behavior over the most recent five rounds.

To compute the trust score of a Player 2 after Round n , we need to first compute the number of times that Player 2 has had an opportunity to make a choice—the

idea being that their trust scores should neither be incremented nor decremented in cases where Player 1 chooses his outside option. This will be recorded as Player 2's *opportunity score*. We need to make a similar allowance to codify the recency effect of trust and trustworthiness. Instead of tracking the behavior of Player 2 (for the purposes of computing her trust score) over the most recent five rounds, we need instead track it over the most recent five rounds *in which she had an opportunity to make a decision*. We will call this queue her *omega queue*.

Algorithm 2 (Player 2 Opportunity Score, Omega Queue). Let c_1 (d_1) indicate a cooperative (defection) move by Player 1, and let c_2 (d_2) indicate a cooperative (defection) move by Player 2. Then Player 2's *opportunity score* in Round n , ρ_n , is given by the following algorithm:

1. If $n = 0$: $\rho_0 = 0$
2. If $n \geq 1$:

$$\rho_n = \begin{cases} \rho_{n-1} & \text{if } d_1 \text{ in Round } n \\ \rho_{n-1} + 1 & \text{if } c_1 \text{ in Round } n \end{cases}$$

Where $n \geq 5$, let Ω_{n-1} be the four most recent rounds prior to Round n in which Player 2 has had a chance to move.

Algorithm 3 (Player 2 Trust Score). Let c_2 (d_2) indicate a cooperative (defection) move by Player 2. Then the trust score of a Player 2 after Round n , τ_n^2 , is given by the following algorithm:

1. If $n = 0$: $\tau_0^2 = 0$
2. If $\rho_n = \rho_{n-1}$: $\tau_n^2 = \tau_{n-1}^2$
3. If $\rho_n \neq \rho_{n-1}$, $\rho_n \leq 5$, and $n \leq 5$: Let k be the number of c_2 moves through Round $n - 1$. Then:

$$\tau_n^2 = \begin{cases} \frac{k}{\rho_n} & \text{if } d_2 \text{ in Round } n \\ \frac{k+1}{\rho_n} & \text{if } c_2 \text{ in Round } n \end{cases}$$

4. If $\rho_n \neq \rho_{n-1}$, $\rho_n \leq 5$, and $n > 5$: Let k be the number of c_2 moves in Ω_{n-1} .

Then:

$$\tau_n^2 = \begin{cases} \frac{k}{\rho_n} & \text{if } d_2 \text{ in Round } n \\ \frac{k+1}{\rho_n} & \text{if } c_2 \text{ in Round } n \end{cases}$$

5. If $\rho_n \neq \rho_{n-1}$ and $\rho_n \geq 5$: Let k be the number of c_2 moves in Ω_{n-1} . Then:

$$\tau_n^2 = \begin{cases} \frac{k}{5} & \text{if } d_2 \text{ in Round } n \\ \frac{k+1}{5} & \text{if } c_2 \text{ in Round } n \end{cases}$$

The two treatments reported below differ according to their *matching protocol*. In the baseline condition (the Random treatment) subjects are *randomly* paired each period. Trust scores in the Random treatment are tracked, but not used in matching Players 1 and Players 2. The experimental treatment (the Sorted treatment) *pairs subjects according to their trust scores*. The matching protocol for the Sorted treatment is straightforward: At the end of Round n Players 1 are rank-ordered by their trust scores (high to low). Similarly for Players 2. Then the matching rule simply pairs the highest ranked Player 1 with the highest ranked Player 2 for interaction in Round $n + 1$, the next to highest ranked Player 1 with the next to highest ranked Player 2 for interaction in Round $n + 1$, and so on.

3.3 Experimental Design and Procedures

We ran four sessions of the Sorted treatment and four sessions of the Random treatment.¹ Each experimental session consisted of 16 subjects.² In all experiments, subjects were paid \$5.00 for showing up on time. At the end of the experiment, their accumulated earnings were paid to them privately (single-blind protocol). The interactions consisted of anonymous pairings in a computerized game. By using a mouse, each Player 1 clicked on the right or down arrows. This information was then displayed on their counterpart's screen. If Player 1 had moved down, Player 2 would be prompted to click on the right or down arrow. This information was then displayed on Player 1's screen. The payoffs represent the experimental dollar amounts the subjects could earn with an exchange rate of 20 experimental dollars equal to 1 U.S. dollar, and are common knowledge. The subjects were undergraduates at the University of Arizona and did not have prior experience with a similar environment.

As a subject enters the Economic Science Laboratory he is paid his show-up fee and immediately (randomly) seated at a computer terminal in a large room containing 40 terminals. Each terminal is in a separate stall, and the subjects are dispersed so that no subject can see the terminal screen of another. Each person is randomly assigned a role (Player 1 or 2) and keep this role for the entirety of the experiment. The games are played sequentially for 20 periods, although the subjects did not know the total number of periods. The experiments lasted on average a little under one hour, from arrival to completion. A subject's total earnings on average is \$27.00 for participating. Each subject participates in one and only one such experiment. See Table 3.1 for a summary of the experimental design.

The instructions differed slightly with respect to information about how the pairings were determined (see Appendix B for detailed instructions). Subjects in the

¹In order to control for some variability we ran all of the sessions at the same time of day, taking two weeks to complete.

²One randomized treatment only had 14 subjects due to no shows.

Sorted	4/64/1280*
Random	4/62/1240

* $a/b/c$ where a = number of sessions, b = number of subjects, c = number of observations.

TABLE 3.1. Experimental Design

random treatment saw the following about matching: “*Each period you will be randomly paired with another individual: your counterpart for that period. You will participate for several periods, being randomly re-paired each period.*” Subjects in the sorted treatment, on the other hand, saw the same statements, but without the word “randomly”: “*Each period you will be paired with another individual: your counterpart for that period. You will participate for several periods. being re-paired each period.*” We did not reveal the exact assignment rule to subjects in the sorted treatment because we were concerned that such information might generate a difference in strategic behavior.³ Knowing that cooperators are being matched each period might lead individuals to alter their type for strategic reasons rather than due to reciprocity type motives.

3.4 Results

Table 3.2 provides the conditional outcome frequencies by blocks of five trials from the Sorted and Random treatments. Note that there are significant differences in the proportion of SPE outcomes across treatments. There is also a significant difference between the proportion of cooperative (25, 25) outcomes being reached under the Sorted matching rule as compared to under the Random matching rule. This is most pronounced in the last trial block. When subjects are sorted based on their trust scores there are far fewer pairs ending up at the SPE; when subjects are sorted.

³Gunthorsdottir, Houser, and McCabe faced similar considerations when using sorting in public goods games.

Trials	(20, 20)	Down	(25, 25)	(15, 30)
Sorted				
1-5	$\frac{74}{160} = .4625$	$\frac{86}{160} = .5375$	$\frac{44}{86} = .5116$	$\frac{42}{86} = .4884$
6-10	$\frac{80}{160} = .50$	$\frac{80}{160} = .50$	$\frac{58}{80} = .725$	$\frac{22}{80} = .2750$
11-15	$\frac{77}{160} = .4813$	$\frac{83}{160} = .5188$	$\frac{70}{83} = .8434$	$\frac{13}{83} = .1566$
16-20	$\frac{75}{160} = .4688$	$\frac{85}{160} = .5313$	$\frac{71}{85} = .8353$	$\frac{14}{85} = .1647$
Random				
1-5	$\frac{73}{150} = .4867$	$\frac{77}{150} = .5133$	$\frac{36}{77} = .4675$	$\frac{41}{77} = .5325$
6-10	$\frac{94}{150} = .6267$	$\frac{56}{150} = .3733$	$\frac{24}{56} = .4286$	$\frac{32}{56} = .5714$
11-15	$\frac{93}{150} = .62$	$\frac{57}{150} = .38$	$\frac{22}{57} = .386$	$\frac{35}{57} = .614$
16-20	$\frac{109}{150} = .7267$	$\frac{41}{150} = .2733$	$\frac{21}{41} = .5122$	$\frac{20}{41} = .4878$

TABLE 3.2. Conditional Outcomes by Trial Block

more pairs reach the cooperative outcome than when they are randomly matched each round. Players 1 reach the SPE 46.88% of the time in the Sorted treatment as compared to 72.67% in the Random treatment. Furthermore, Players 2 who are paired with trusting Players 1 respond in kind in the Sorted treatment 83.53% of the time compared with 51.22% of the time in the Random treatment.⁴

The above is aggregated in trial blocks. The dynamics of play over time reveals the same trends, albeit more graphically. Figures 3.2-3.5 show the mean fraction of each type of play over the 20 rounds for both treatment conditions. The trends are unmistakable: as play proceeds through the later rounds, cooperation emerges and is sustained among the sorted subjects, but there is no similar round-effect for the

⁴It was interesting watching the results come in from these experiments. What was easy to observe is that by Round 10 in the Sorted treatment around half of the Players 1 were playing SPE. so their trust scores began deteriorating rapidly and about half were playing down, keeping their trust scores near the maximum. Almost all of the trusting interactions were met with trustworthiness by their counterpart, keeping half of the Players 2 trust scores high as well. This was not the case in the Random treatment.

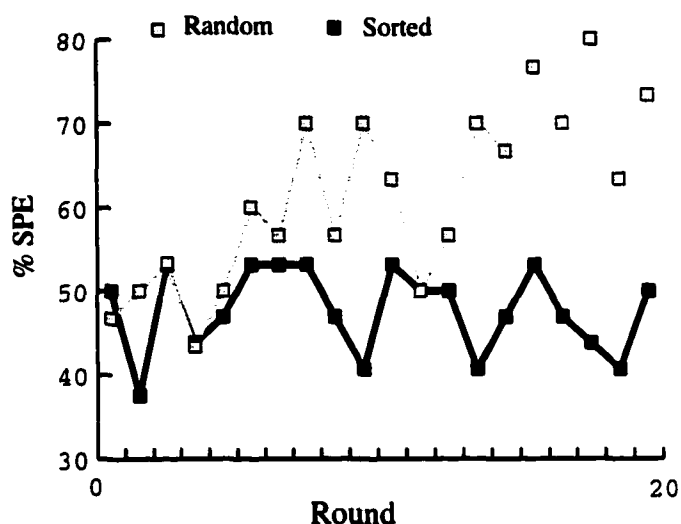


FIGURE 3.2. Percent of Players 1 Choosing SPE Over Time

randomly paired subjects.

Along these same lines, it is interesting to look at the average trust score by blocks of five trials (See Table 3.3). Remember that in both the Random and Sorted treatment a trust score is calculated for each player based on their decisions, but only the Sorted treatment matches players according to their score. Since the trust scores track the behavioral data, it is not surprising that an examination of the scores tells a very similar story to that of the outcome frequencies. The average trust score over the first 10 trials is statistically the same for the two treatments (p -value = .5246 for 1–5 and p -value = .1331 for 6–10). However, in the last 10 trials the scores are significantly higher under the Sorted condition than in the Random (p -value = .0008 for 11–15 and p -value = .0000 for 16–20). This is recorded in Table 3.3.

That mean trust scores are higher in the Sorted treatment does not decide between two competing hypotheses. Is it the case that the sorting mechanism induces a high degree of trust and reciprocity among pairs or is it merely the case that first movers are extremely trusting in a population of defecting Players 2? In every interaction in this environment there is more than monetary costs and benefits at stake: there

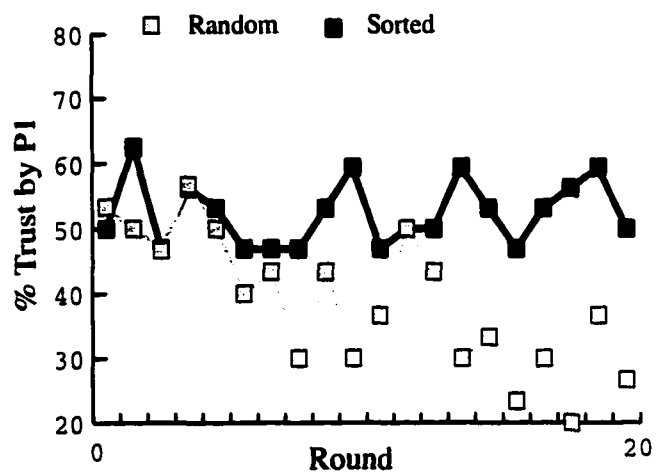


FIGURE 3.3. Percent of Players 1 Trusting Over Time

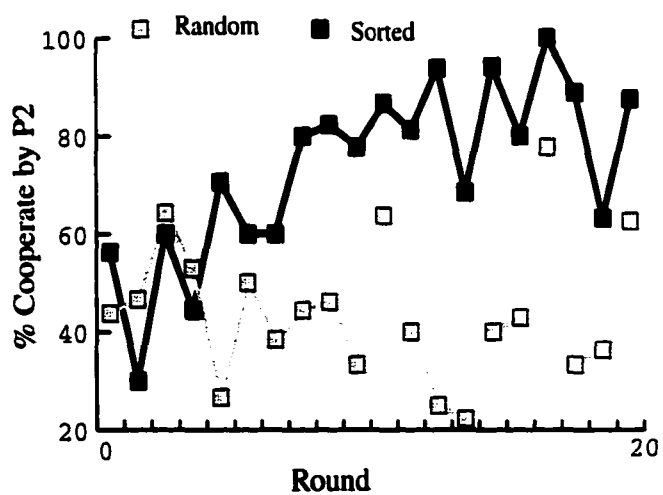


FIGURE 3.4. Percent of Players 2 Cooperating Over Time

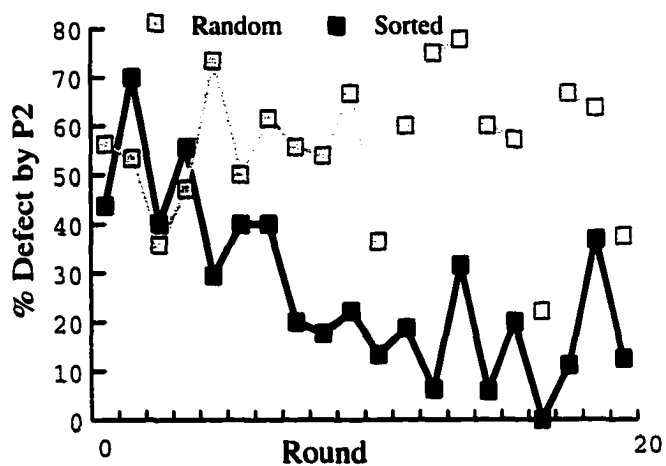


FIGURE 3.5. Percent of Players 2 Defecting Over Time

Trials	1-5	6-10	11-15	16-20
Mean (Sorted)	0.45	0.48	0.50	0.52
Mean (Random)	0.43	0.44	0.40	0.36
Stdev (Sorted)	0.4030	0.3787	0.4337	0.4406
Stdev (Random)	0.4165	0.3659	0.3531	0.3472

TABLE 3.3. Trust Score by Condition

are potential social costs and benefits as well. There are two ways for players to incur social costs. If they defect then their trust scores are decremented; if they are defected upon then they incur the cost of being trusting when they ought not have been. Similarly, there are two ways to incur social gain. One is through making a cooperative choice; the other is when players actually reach the cooperative outcome. We would like to track how efficient choices are with respect to these potential social gains. We can introduce an *efficiency score* at round n for each player, $\nu_n^i \in [0, 1]$, as follows:

$$\nu_n^i = \frac{\tau_n^i + d}{2}$$

where $d = 0$ if Player i did not reach the cooperative outcome in Round n and $d = 1$ if she did.

Table 3.4 displays some descriptive statistics for the variable, efficiency score, under each condition in five trial blocks. Figure 3.6 plots the average efficiency score for both treatments at each round. The efficiency scores begin in Round 1 at less than 0.40 for both treatments and remain similar in magnitude through, roughly, the first nine rounds. However, in the later rounds, the efficiency being achieved in each condition is significantly different. In fact, the null hypothesis that there is no difference in the average efficiency score of the two treatments can be easily rejected (see Table 3.5; p -value = .0000). The level of efficiency being achieved is greater when subjects are being matched according to their trust score.

We can classify subjects as either a non-cooperator or cooperator based on their first observed move. So Players 1 are a non-cooperating type if in Round 1 they chose (20, 20) and a cooperating type if they chose to play down, passing the game to their counterpart. Similarly, for Players 2. A Player 2 is a non-cooperating type if when her counterpart first played down, she chose the defection outcome (15, 30), and a Player 2 is a cooperating type if she chose the cooperative outcome (25, 25)

Trials	1-5	6-10	11-15	16-20
Mean (Sorted)	0.3624	0.4219	0.4698	0.4838
Mean (Random)	0.3344	0.2981	0.2738	0.2493
Mode (Sorted)	0	0	1	1
Mode (Random)	0	0	0	0
Stdev (Sorted)	0.3802	0.3968	0.4507	0.4501
Stdev (Random)	0.3725	0.3069	0.2911	0.2952

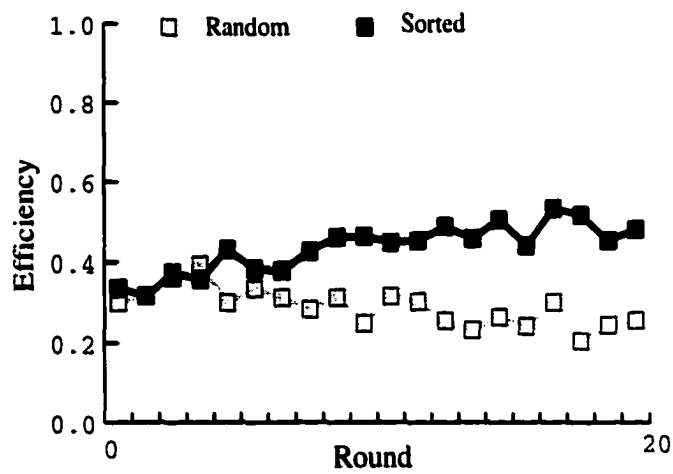
TABLE 3.4. Efficiency Score, ν , by Condition

FIGURE 3.6. Average Efficiency Scores Over Time

Source of Variation	SS	DF	MS
Between	13.1225	1	13.1225
Within	350.19	2478	0.1416
Total about Grand Average	364.0415	2479	

TABLE 3.5. ANOVA for Efficiency Score ν

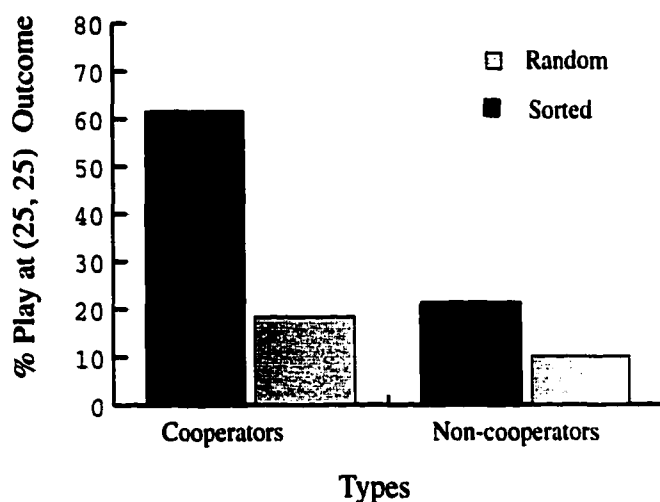


FIGURE 3.7. Cooperators versus Non-cooperators: Percent of Each Type Reaching the Cooperative Outcome of (25, 25)

on her first available move. Once we establish this typing, we can analyze how play differs among these groups depending on whether they are being sorted by their trust scores or simply being randomly re-paired. We want to focus on the last 10 trials in particular (see Figure 3.7). Cooperators fare much better when they are meeting other cooperators under the sorting mechanism than when they randomly meet their counterparts—the last 10 interactions result in an outcome of (25, 25) 62% of the time in the Sorted treatment as compared to only 18% of the time in the Random treatment (p -value = .0000). This is not the case for non-cooperative types. In fact, there is no treatment effect for the defecting types: the percentage of cooperative outcomes reached in the last 10 trials is not statistically different between the Random and Sorted treatments (p -value = .1187).

In summary form, here are the five central results from these sorting experiments.

Result 1. In the last 10 rounds, the fraction of subjects reaching the SPE is dramatically lower in the Sorted treatment than in the Random treatment.

Result 2. In the last 10 rounds, the fraction of subjects reaching the cooperative

outcome is significantly higher in the Sorted treatment than in the Random treatment.

Result 3. In the last 10 rounds, the average trust scores are much higher in the Sorted treatment than in the Random treatment.

Result 4. The average efficiency score, i.e. how efficient play is with respect to the potential social benefit, is higher in the Sorted treatment than in the Random treatment.

Result 5. In the last 10 rounds the number of cooperative player types reaching cooperative outcomes is far greater in the Sorted treatment than in the Random treatment. There is no treatment effect for non-cooperative types.

3.5 Conclusion

It is well-known in evolutionary game theory that population clustering in PD games allows for some cooperative strategies to invade populations of stable defecting strategies. Similarly, in the experimental community there are results which suggest that a similar “clustering” phenomenon can be induced among subjects in public goods games to sustain high levels of contributions. The results of the sorting experiments here suggest a similar story about behavior in simple two-person bargaining games. Sorting subjects by trust scores accomplishes two tasks. First, it allows cooperative play which is Pareto-superior to that predicted by non-cooperative game theory to emerge. Second, once cooperative play emerges, sorting subjects does not allow this behavior to be “infected” and compromised by either defecting Players 2 or by untrusting Players 1.

Appendix A
INSTRUCTIONS FOR VOLUNTARY/INVOLUNTARY
EXPERIMENTS

Page 1

In this experiment you will participate in a two person decision problem. You will participate in the decision problem once. You will be randomly paired with another individual: your counterpart. The joint decisions made by you and your counterpart will determine how much money you will earn.

Your earnings will be paid to you in cash at the end of the experiment. We will not tell anyone else your earnings. We ask that you do not discuss your earnings with anyone else.

Please read the following instructions carefully. If you have a question at any time, please raise your hand and someone will come by to help.

Page 2

You will see a diagram similar to this one at the beginning of the experiment. You and another person will participate in a decision problem like the diagram below. We will refer to this other person as your counterpart.

SCREEN DIAGRAM

You and your counterpart will be either DM 1 or DM 2. Beside the diagram we show this information. Right now you are DM 1. Please click "Next" to continue.

Page 3

Notice the boxes with letters in them. These letters will be replaced by numbers representing DOLLAR AMOUNTS during the experiment. The boxes with numbers

show the different earnings that you and your counterpart can make. There are two numbers in each box. The number on the top (which is indented now) is DM 1's earnings if this box is reached. The number on the bottom is DM 2's earnings.

SCREEN DIAGRAM

You and your counterpart will jointly determine a path through the diagram to an earnings box. Please click "Next" to continue.

Page 4

A path is defined as sequence of moves through the diagram.

A move is a choice of direction in the diagram.

PATH STARTS HERE SCREEN DIAGRAM

The arrows in the diagram show the possible directions of moves that can be made. Notice that the moves for both DM 1 and DM 2 are always DOWN or RIGHT. When you click on either arrow, the path is highlighted.

The circles in the diagram with numbers in them indicate who gets to move at that point in the diagram. Please click "Next" to continue.

Page 5

For example, DM 1 starts the process at the top of the diagram by moving right or down. If DM 1 moves right, DM 2 will have a decision to make. We'll show you what this looks like later.

DIAGRAM WITH ALL ARROWS SHOWING MOVES

If DM 1 moves down, it is DM 2's turn to move. DM 2 can then move down or right. If DM 2 moves right, DM 1 earns 'wig' and DM 2 earns 'wog'. If DM 2 moves down, DM 1 earns 'xig' and DM 2 earns 'xog'.

The decision path that was chosen will be highlighted. Please click "Next" to continue.

Page 6

We will now show you what the decisions look like from the point of view of DM 1. When you are DM 1 you move first. The arrows show you can move right or down. In order to move, click on the arrow for your choice. DM 2 will only see your decision when you click the "Send" button to finalize your decision. To see how this works, click the RIGHT ARROW now. Be sure to click "Send" to finalize your move.

DIAGRAM ALLOWING MOVE RIGHT

At this point the moves are over. The path taken is highlighted white and earnings received are highlighted. Please click 'Next' to continue.

Page 7

As another example as DM 1, move DOWN by clicking on the arrow. To confirm your move click the "Send" button.

DIAGRAM TO MOVE DOWN

Once the subject makes the choice, the following appears: Since you moved Down as DM 1, DM 2, seeing your move, now has a decision to make. If DM 2 moves right then you would earn 'wig' and DM 2 would earn 'wog'. If DM 2 moves down then you would earn 'xig' and DM 2 would earn 'xog'. Please click "Next" to continue.

Page 8

We will now show you what decisions look like from DM 2's point of view. Notice that your earnings are indented and this is the BOTTOM NUMBER in the boxes. You will only have a move if DM 1 moves down. Suppose DM 1 has moved down. You have to decide to move right or down. Please make a choice now by clicking on the arrow of your choice. Then click "Send" to confirm your move.

SCREEN DIAGRAM

Either the subject moves Right as DM 2 in which case she sees the following: Since you moved Right as DM 2, DM 1's earnings are 'wig'. Your earnings are 'wog'. Please click "Next" to continue.

OR the subject moves Down as DM 2 in which case she sees the following: Since you moved Down as DM 2, DM 1's earnings are 'xig'. Your earnings are 'xog'. Please click "Next" to continue.

Page 9

IMPORTANT POINTS:

- * You will be randomly paired with another individual: your counterpart.
- * You will participate in the decision problem once.
- * If you are DM 1 then your counterpart will be DM 2. In this case, you will make a decision first. On the other hand, if you are DM 2, your counterpart will be DM 1. If this is the case, you will have a decision to make if DM 1 chooses down.
- * If you are DM 1, your payoff is the top number in the box. If you are DM 2, your payoff is the bottom number in the box. You will receive that amount of money if the box is reached. The numbers represent dollar amounts.

This concludes the directions. If you wish to return to them please click the "Back" button. If you have any questions please raise your hand. Otherwise, to begin the experiment, please click the green button, "Finished with directions".

Appendix B

INSTRUCTIONS FOR SORTING EXPERIMENTS

Page 1

In this experiment you will participate in a series of two person decision problems. The experiment will last for several periods. Each period you will be randomly paired with another individual: your counterpart for that period. The joint decisions made by you and your counterpart for that period will determine how much money you will earn in that period. After each period you will be randomly re-paired.

Your earnings will be paid to you in cash at the end of the experiment. We will not tell anyone else your earnings. We ask that you do not discuss your earnings with anyone else.

Please read the following instructions carefully. If you have a question at any time, please raise your hand and someone will come by to help.

Page 2

Notice that another button, "Back", has appeared at the bottom of the page. If at any time you wish to return to a previous page, click "Back". To continue reading the directions, click "Next".

Page 3

You will see a diagram similar to this one at the beginning of the experiment. You and another person will participate in a decision problem like the diagram below. We will refer to this other person as your counterpart.

SCREEN DIAGRAM

One of you will be DM 1. The other person will be DM 2. Beside the diagram we

show whether you are DM 1 or DM 2. In this example, for now, you are DM 1. Please click "Next" to continue.

Page 4

Notice the boxes with letters in them. These letters will be replaced by numbers representing Experimental Dollars during the experiment. For 20 Experimental Dollars you will earn 1 U.S. dollar. The boxes with numbers show the different earnings in Experimental Dollars that you and your counterpart can make. There are two numbers in each box. The number on the top (which is indented now) is DM 1's earnings if this box is reached. The number on the bottom is DM 2's earnings.

SCREEN DIAGRAM

You and your counterpart will jointly determine a path through the diagram to an earnings box. Please click "Next" to continue.

Page 5

A path is defined as sequence of moves through the diagram.

A move is a choice of direction in the diagram.

SCREEN DIAGRAM

The arrows in the diagram show the possible directions of moves that can be made. Notice that the moves for both DM 1 and DM 2 are always DOWN or RIGHT. When you click on either arrow, the path is highlighted.

The circles in the diagram with numbers in them indicate who gets to move at that point in the diagram. Please click "Next" to continue.

Page 6

For example, DM 1 starts the process at the top of the diagram by moving right or down. If DM 1 moves right the experiment is over. DM 1 earns 'zig' and DM 2 earns

'zog'.

SCREEN DIAGRAM

If DM 1 moves down, it is DM 2's turn to move. DM 2 can move right or down. If DM 2 moves right, DM 1 earns 'wig' and DM 2 earns 'wog'. If DM 2 moves down, DM 1 earns 'xig' and DM 2 earns 'xog'.

The decision path that was chosen will be highlighted. Please click "Next" to continue.

Page 7

We will now show you what the decisions look like from the point of view of DM 1. When you are DM 1 you move first. The arrows show you can move right or down. In order to move, click on the arrow for your choice. DM 2 will only see your decision when you click the "Send" button to finalize your decision. To see how this works, click the RIGHT ARROW now. Be sure to click "Send" to finalize your move.

SCREEN DIAGRAM

At this point the moves are over. The path taken is highlighted white and earnings received are highlighted. Please click 'Next' to continue.

Page 8

As another example as DM 1, move DOWN by clicking on the arrow. To confirm your move click the "Send" button.

SCREEN DIAGRAM

Once the subject makes the choice, the following appears: Since you moved Down as DM 1, DM 2, seeing your move, now has a decision to make. If DM 2 moves right then you would earn 'wig' and DM 2 would earn 'wog'. If DM 2 moves down then you would earn 'xig' and DM 2 would earn 'xog'. Please click Next to continue.

Page 9

We will now show you what decisions look like from DM 2's point of view. Notice that your earnings are indented and this is the **BOTTOM NUMBER** in the boxes. You will only have a move if DM 1 moves down. Suppose DM 1 has moved down. You have to decide to move right or down. Please make a choice now by clicking on the arrow of your choice. Then click "Send" to confirm your move.

SCREEN DIAGRAM

Either the subject moves Right as DM 2 in which case she sees the following: Since you moved Right as DM 2, DM 1's earnings are 'wig'. Your earnings are 'wog'. Please click "Next" to continue.

OR the subject moves Down as DM 2 in which case she sees the following: Since you moved Down as DM 2, DM 1's earnings are 'xig'. Your earnings are 'xog'. Please click "Next" to continue.

Page 10**IMPORTANT POINTS:**

- * Each period you will be randomly paired with another individual: your counterpart for that period.
- * You will participate for several periods, being randomly re-paired each period.
- * If you are DM 1, your counterpart will be DM 2. In this case, you will make a decision first. On the other hand, if you are DM 2, your counterpart will be DM 1. If this is the case, you will have a decision to make if DM 1 chooses down.
- * If you are DM 1, your payoff in Experimental Dollars is the top number in the box. If you are DM 2, your payoff in Experimental Dollars is the bottom number in the box. You will receive that amount of money if the box is reached. For every 20 Experimental Dollars you earn, you will receive 1 U.S. Dollar.

This concludes the directions. If you wish to return to them please click the "Back" button. If you have any questions please raise your hand. Otherwise, to begin the experiment, please click the green button, "Finished with directions".

REFERENCES

- Abbink, Klaus, Bernd Irlenbusch, and Elke Renner (2000). "The Moonlighting Game: An Experimental Study on Reciprocity and Retribution." *Journal of Economic Behavior and Organization* 42(2): 265–277.
- Axelrod, Robert (1984). *The Evolution of Cooperation*. New York: Basic Books.
- Axelrod, Robert (1997). *The Complexity of Cooperation*. Princeton: Princeton University Press.
- Baron-Cohen, Simon (1995). *Mindblindness*. Cambridge, MA: MIT Press.
- Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). "Trust, Reciprocity, and Social History," *Games and Economic Behavior* 10: 122–142.
- Bolton, Gary and Axel Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity, and Competition" *American Economic Review* 90(1): 166–193.
- Binmore, Kenneth (1999). "Why Experiment in Economics?" *The Economic Journal* 109: 16–24.
- Buchan, Nancy R., Eric J. Johnson, and Rachel T.A. Croson (2000). "Trust and Reciprocity: An International Experiment," Working Paper.
- Burnham, Terry, Kevin McCabe, and Vernon L. Smith (2000). "Friend-or-foe Intentionality Priming in an Extensive Form Trust Game." *Journal of Economic Behavior and Organization* 1244: 1–17.
- Camerer, Colin F. (1997). "Progress in Behavioral Game Theory," *Journal of Economic Perspectives* 11(4): 167–188.
- Carruthers, Peter and Peter K. Smith (1996). *Theories of Theories of Mind*. New York, NY: Cambridge University Press.
- Charness, Gary and Matthew Rabin (2000). "Social Preferences: Some Simple Tests and a New Model," Working Paper.
- Cosmides, Leda and John Tooby (1992). "Cognitive Adaptations for Social Exchange," in Barkow, Cosmides, and Tooby (eds.), *The Adapted Mind* (New York: Oxford University Press).
- Croson, Rachel and Nancy Buchan (1999). "Gender and Culture: International Experimental Evidence from Trust Games," *American Economic Review* 89(2): 386–391.

- Dufwenberg, Martin and Uri Gneezy (2000). "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior* 30(2): 163–182.
- Dufwenberg, Martin and Georg Kirchsteiger (1998). "A Theory of Sequential Reciprocity," Tilburg CentER for Economic Research Discussion Paper: 9837.
- Eckel, Catherine C. and Philip J. Grossman (2001). "Chivalry and Solidarity in Ultimatum Games," *The Economic Journal* 39(2): 171–188.
- Eckel, Catherine C. and Philip J. Grossman (1998). "Are Women Less Selfish Than Men?: Evidence from Dictator Experiments," *The Economic Journal* 108: 726–735.
- Eckel, Catherine C. and Philip J. Grossman (1996). "Altruism in Anonymous Dictator Games," *Games and Economic Behavior* 16(2): 181–191.
- Falk, Armin, Ernest Fehr, and Urs Fischbacher (1999). "On the Nature of Fair Behavior," Institute for Empirical Research in Economics, The University of Zurich. Working Paper No.17.
- Falk, Armin, and Urs Fischbacher (1999). "A Theory of Reciprocity." Institute for Empirical Research in Economics, The University of Zurich. Working Paper No.6.
- Fehr, Ernst and Simon Gächter (2000). "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, forthcoming.
- Fehr, Ernst and Klaus M. Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics* 114(3): 817–868.
- Forsythe, Robert, Joel Horwitz, N.E. Savin, and Martin Sefton (1994). "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior* 6: 347–369.
- Fudenberg, Drew and Edward Maskin (1990). "Evolution and Cooperation in Noisy Repeated Games," *American Economic Review* 80: 274–279.
- Fudenberg, Drew and Jean Tirole (1991). *Game Theory*. Cambridge, MA: MIT Press.
- Gillies, Anthony S. and Mary L. Rigdon (2000). "Folk-psychology and Strategic Interaction: Varieties of Mind-Reading in Two-Person Trust Games," Working Paper, University of Arizona.
- Gunnthorsdottir, Anna, Dan Houser, and Kevin McCabe (2000). "Excluding Free-riders Improves Reciprocity and Promotes the Private Provision of Public Goods." Working Paper, The University of Arizona.

- Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982). "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* **3**: 367–388.
- Güth, Werner, Steffen Huck, and Peter Ockenfels (1996). "Two-level Ultimatum Bargaining with Incomplete Information," *The Economic Journal* **106**: 593–604.
- Güth, Werner and Eric van Damme (1998). "Information, Strategic Behavior, and Fairness in Ultimatum Bargaining: An Experimental Study." *Journal of Mathematical Psychology* **42**: 227–247.
- Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith (1994). "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* **7**: 346–380.
- Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith (1996). "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review* **86**: 653–660.
- Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith (1998). "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology." *Economic Inquiry* **36**: 335–352.
- Kagel, John H. and Katherine Wolfe (2000). "Tests of Difference Aversion to Explain Anomalies in Simple Bargaining Games," Working Paper.
- Kritikos, Alexander and Friedel Bolle (1999). "Approaching Fair Behavior." Working Paper No. 143, Europa-Universität Viadrina.
- Langer, Elaine (1975). "The Illusion of Control," *Journal of Personality and Social Psychology* **32**: 311–328.
- Ledyard, John O. (1995). "Public Goods," in Kagel and Roth (eds.), *The Handbook for Experimental Economics* (Princeton: Princeton University Press).
- Levine, David K. (1998). "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* **1**(3): 593–622.
- McCabe, Kevin, Stephen Rassenti, and Vernon Smith (1996). "Game Theory and Reciprocity in Some Extensive Form Experimental Games." *Proceeding of the National Academy of Sciences* **93**: 13421–13428.
- McCabe, Kevin, Stephen Rassenti, and Vernon Smith (1998). "Reciprocity, Trust, and Payoff Privacy in Extensive Form Bargaining," *Games and Economic Behavior* **24**: 10–24.

- McCabe, Kevin, Mary Rigdon, and Vernon Smith (2001). "Cooperation in Single Play. Two-Person Extensive Form Games between Anonymously Matched Players." in R. Zwick (ed.), *Experimental Business Research* (Boston, MA: Kluwer).
- McCabe, Kevin and Vernon Smith (2000). "A Comparison of Naïve and Sophisticated Subject Behavior with Game Theoretic Predictions," *Proceedings of the National Academy of Sciences* **97**(7): 3777-3781.
- McCabe, Kevin and Vernon Smith (2000). "Goodwill Accounting in Economic Exchange," in Gerd Gigerenzer and Reinhard Selten (eds.). *Bounded Rationality: The Adaptive Toolbox* (Cambridge, MA: MIT Press).
- McCabe, Kevin, Vernon Smith, and Michael LePore (2000). "Intentionality Detection and 'Mindreading': Why Does Game Form Matter?" *Proceeding of the National Academy of the Sciences* **97**(8): 4404-4409.
- Nash, John F., Jr. (1950). "The Bargaining Problem," *Econometrica* **18**: 155-162. Reprinted in: H. W. Kuhn (ed.), *Classics in Game Theory* (Princeton: Princeton University Press, 1997).
- Orbell, John M. and Robyn M. Dawes (1993). "Social Welfare. Cooperators' Advantage, and the Option of Not Playing the Game," *American Sociological Review* **58**: 787-800.
- Rabin, Matthew (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review* **83**(5): 1281-1302.
- Rapoport, Amnon (1997). "Order of Play in Strategically Equivalent Games in Extensive Form," *International Journal of Game Theory* **26**(1): 113-136.
- Roth, Alvin (1995). "Bargaining Experiments," in Kagel and Roth (eds.). *The Handbook for Experimental Economics* (Princeton: Princeton University Press).
- Schotter, Andrew, Avi Weiss, and Inigo Zapater (1996). "Fairness and Survival in Ultimatum and Dictatorship Games," *Journal of Economic Behavior and Organization* **31**(1): 37-56.
- Smith, Vernon L. (2001). "Experimental Methods in (Neuro)Economics," *Encyclopedia of Cognitive Science* (forthcoming).
- Smith, Vernon L. (1998). "The Two Faces of Adam Smith," *Southern Economic Journal* **65**(1): 1-19.
- Smith, Vernon L. (1964). "Effect of Market Organization on Competitive Equilibrium," *Quarterly Journal of Economics* **78**:181-201.

- Van der Heijden, Eline C. M., Jan H.M. Nelissen, Jan J.M. Potters. and Harrie A.A. Verbon (1998). "The Poverty Game and the Pension Game: The Role of Reciprocity," *Journal of Economic Psychology* 19:5-41.
- de Waal, Frans (1997). "Food Transfers Through Mesh in Brown Capuchins." *Journal of Cognitive Psychology* 111(4): 370-378.
- de Waal, Frans (1996). *Goodnatured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge: Harvard University Press.
- Weibull, Jörgen W. (1995). *Evolutionary Game Theory*. Cambridge: MIT Press.