Trust, Fear, Reciprocity, and Altruism: *Theory and Experiment*

By James C. Cox*

This paper describes central topics in our research program on social preferences. The discussion covers experimental designs that discriminate among alternative components of preferences such as unconditional altruism, positive reciprocity, trust (in positive reciprocity), negative reciprocity, and fear (of negative reciprocity). The paper describes experimental data on effects of social distance and decision context on reciprocal behavior and male vs. female and group vs. individual differences in reciprocity. The exposition includes experimental designs that provide direct tests of alternative models of social preferences and summarizes implications of data for the models. The discussion reviews models of other-regarding preferences that are and are not conditional on others' revealed intentions and the implications of data for these models.

1. Introduction

The paper describes our research program on social preferences in which the central objective is to improve theory through a program of experimental testing and theoretical modeling motivated by data. Content will be drawn from several papers, and co-authors will be cited in the context of questions addressed by the research.

There are many other research programs in this area but I will only be discussing my own and my co-authors' research (with apologies to other researchers whose work is not discussed). The focus on our own research program facilitates a structured discussion of the relationship between specific features of experimental designs and theory development objectives.

This research program on social preferences is based on acceptance of the objective of parsimony in theoretical modeling (Samuelson, 1947), of never including within a model any complication that is not necessary to explain the phenomena being studied. Application of parsimony to theoretical modeling of social preferences and design of experiments to test the models is interpreted as leading to a focus first on the question of when the "economic man" model does not predict well, and models of "other-regarding" preferences are needed, and subsequently on when the other-regarding preferences need to include beliefs and/or intentions.

The logic of this application of parsimony is as follows. We begin by noting that the simplest model one can develop is the "economic man" model of self-regarding preferences in which the only thing an agent cares about in any context is his own material rewards. As is well documented by a large literature, the self-regarding preferences model does not predict well in many contexts in which distributional fairness is a salient concern.[1] This suggests that we first consider models of unconditional other-regarding preferences because such distributional preferences can be introduced into economic models by simply redefining the goods over which preferences are defined while preserving conventional regularity properties of the models such as completeness, transitivity, convexity and, perhaps, monotonicity. But if the other-regarding preferences that are modeled are conditional on – or "include" – reactions to others' past actions or beliefs about their future actions then that is a more fundamental departure from traditional economic theory. And so again, by the parsimony objective of theoretical modeling, one does not want to introduce such complications into models if, or when, they are not needed to maintain consistency between theory and data.

The approach of our research program is based on experimental designs that discriminate between the observable implications of unconditional other-regarding preferences and conditional preferences involving reactions to others' prior actions (such as positive or negative reciprocity) or beliefs about others' future actions (such as trust or fear). The reason to make these discriminations is that unconditional other-regarding preferences can be modeled without introducing intentions or beliefs by simply expanding the identity of goods to include other agents' incomes or consumption goods. In contrast, reciprocity makes preferences over goods dependent on perceptions of others' past actions (or attributions of their intentions) and trust or fear makes preferences over goods dependent on beliefs about others future reactions to one's own actions.

Another manifestation of application of the parsimony objective is that if experiments reveal that one needs to incorporate intentions and beliefs in some contexts, but not others, then it

is desirable to develop a unified approach to modeling behavior in games both with and without reciprocal motivation. I shall discuss some models in the literature and our direct tests of those models, and then review some new models that my co-authors and I are developing for distinct patterns of behavior that are conditional, or are *not* conditional, on others' revealed intentions. When discussing these models, I shall explain how they are variations on the same underlying model and hence an example of this last application of the parsimony objective, of developing a unified approach to modeling both less and more complicated instances of social preferences.

I begin with a game that provides an illustration of our approach to experimenting with fairness games. The game used in the discussion is the investment game.

2. An Example: The Investment Game

An experiment with the investment game has the following characteristics (see, for examples, Berg, et al., 1995; Cox, 2004). Subjects are randomly paired. Each subject in each pair is given $10. Second movers are told to keep their $10. A first mover can either keep her $10 or give some or all of it to the second mover. Any amount given is multiplied by 3 by the experimenter. A second mover can either keep all of any amount received or return part or all of it to the paired first mover. The game is played only once. The experimental protocol uses double blind payoffs in which any individual subject's responses are anonymous to the experimenter and other subjects. All of the features of the experimental design and protocol are common information given to all subjects.

2.1 Predictions of the "Economic Man" Model

Predictions of the traditional economic man model for this game are transparent. Since second movers care only about their own material gain, they will keep any tripled amount sent by first movers. Since first movers care only about their own material gain, and know that second movers have the same kind of preferences, first movers will send nothing. Zero amounts returned

and sent are the subgame perfect equilibrium of this game, given the economic man assumption about preferences. The predicted outcome is inefficient: each subject pair is predicted to get $20 in payoff – only the endowment – when it could have gotten as much as $40.

2.2 Behavior

Experiments with the investment game have been conducted by several researchers and the results look pretty much the same regardless of who runs the experiment. Figure 1 shows behavior in the investment game reported in Cox (2004). The amounts sent are represented by the striped bars and the amounts returned are portrayed by the solid bars. Of course, what the economic man model predicts is that there won't be any visible bars (of either type) in Figure 1. There are six subject pairs shown at the left side of the figure with no bars. The other 26 subject pairs do not behave like that. The overwhelming majority of first movers send money. Some second movers who receive money keep it all, as the economic man model predicts. So there are a few cases in which the first movers did *not* behave like economic man and the second movers did. But a large proportion of the second movers did not behave according to the economic man model either. There are even four subject pairs in which the first movers sent all $10 and the second movers returned $20, in other words the second movers chose the equal split fairness focal point in which each subject in the pair gets $20, exactly double his/her endowment.

Behavior in the investment game is representative of many games in the literature in which deviations from the economic man model's predictions are consistent with trust (by first movers) and reciprocity (by second movers). And many authors have concluded that trust and reciprocity has been observed in the experimental games like this. But the experimental design actually does not support that conclusion. The reason is that first movers may send money to second movers because of unconditional altruistic preferences: it only costs a second mover 33 cents for each $1 increase in the other person's money payoff. Furthermore, second movers may

return money to first movers (who have less, after sending some of their endowment) because of either unconditional altruism or inequality aversion.[2]

If this behavior could all be explained by unconditional altruism then that would be a relatively parsimonious extension of theory: just define the preferences over both my income and your income, assume positive monotonicity in both variables, and assume convexity of indifference curves. On the other hand, if subjects' behavior *is* characterized by trust and/or reciprocity then the implied changes in theory are less parsimonious. Modeling trust requires introduction of beliefs into theory. Modeling reciprocity requires introduction of perceived intentions into theory. These are more extensive and less tractable changes in theory than is modeling unconditional altruism and – according to the objective of parsimony – one does not want to introduce these complications into theory if they are not needed to explain behavior.

In order to proceed without ambiguity in discussing the relation between theory and alternative experimental designs, one needs some clearly-stated definitions of terms. Here are ones that will be used.

2.3 Definitions of Terms for Identifying Behavioral Motivations

Self-regarding (or "economic man") preferences are characterized by positively monotonic utility for one's own material payoffs and indifference about others' material payoffs. *Other-regarding* preferences are characterized by utility that is not constant with respect to variations in one's own or others' material payoffs. *Altruistic* preferences are characterized by utility that is monotonically increasing in others' material payoffs as well as one's own payoffs. *Positive (direct) reciprocity* is a motivation to adopt a generous action that benefits someone else, at one's own material cost, because that person's intentional behavior was perceived to be beneficial to oneself. *Trust* is a belief that one agent has about another. A trusting action is one that creates the possibility of mutual benefit and the risk of loss of one's own utility if the other person defects.

Why did I use the term "utility" in the definition of trust? Because if one is an altruist and would like to send some money to the other person, even if it was certain they wouldn't send anything back, then if the person doesn't, in fact, send anything back there may be no loss in utility. In that case, "trust" wouldn't be needed to explain the first mover's behavior; instead, the more parsimonious explanation – unconditional altruism – would suffice. Thus the question about identifying trusting behavior becomes: "Does a first mover in the investment game send *more* money to the second mover than he would in another game in which the first mover has the same set of feasible choices as in the investment game but knows for sure that the second mover cannot return anything?" This is clearly a different question than: "Does the first mover send any money to the second mover in the investment game?"

Negative (direct) *reciprocity* is a motivation to adopt an action that harms someone else, at one's own material cost, because that person's intentional behavior was perceived to be harmful to oneself. *Fear* is a belief that one agent has about another. An action that is fearful of another is one that forgoes an otherwise preferred action because of a belief that the other agent will inflict costly punishment as a response to choice of the otherwise-preferred action. Negative reciprocity and fear will be discussed in section 3, in the context of experiments with the moonlighting game, but we first continue the discussion of experiments with the investment game.

2.4 Investment Game Triadic Design

Consider again the investment game, but instead of running it by itself, embed it in an experiment with three games, in what we call a "triadic design." Each game is an experimental treatment. The objective of the triadic experimental design is to construct treatments that reveal whether behavior in a central game of interest (in the present case, the investment game) can be represented with a model of unconditional other-regarding preferences or, instead, requires the less parsimonious approach of constructing a model that incorporates agents' attribution of the

intentions revealed by others' past actions and/or beliefs about their future actions. In order to support observational discrimination between these distinct motives for behavior, dictator control treatments are designed so as to provide subjects with the same (own income, other's income) feasible choice sets as does the investment game but remove the decision opportunity of the paired subject, and thereby remove the possible effects of beliefs and intentions attribution on behavior.

The experimental design for the investment game includes the following three treatments. Treatment A is the investment game in which each first mover and each second mover is given a $10 endowment and each $1 increase in the second mover's money payoff costs the first mover 33 cents. Treatment B is a dictator game, with the same endowments as the investment game, which gives dictators the same feasible set of choices (over the ordered pairs of their own and the other's payoffs) that first movers have in the investment game. So what is the difference? First movers have exactly the same decisions to make, and the same feasible set, in treatment A and treatment B. The difference is that they know for sure that in treatment B the second movers cannot return anything. So if we observe that subjects send significantly less in treatment B than they do in treatment A, then we can conclude that amounts sent in treatment A cannot be fully explained by altruism, that we need something else, and the natural thing of course is trust. Why? Because in treatment A the first movers can trust that the second movers will share part of the increased total payoff from the tripling of amounts sent, and as a result send more in treatment A than in treatment B.

Treatment C is the dictator control treatment for reciprocity. Treatment C gives dictators the same choices and feasible sets that second movers have in the investment game. Treatment C is constructed as follows. The dictators correspond to the second movers in the investment game (treatment A). Of course, the non-dictators do not have a decision to make. Each dictator is given a $10 endowment. Each non-dictator is given an endowment equal to the amount kept (i.e. *not* sent) by a specific first mover in treatment A. Furthermore, each dictator is given an additional

dollar amount equal to the amount received by a specific second mover in treatment A from the tripled amount sent by a first mover in treatment A. The subjects are informed with a table of the exact inverse relation between the number of additional dollars received by a dictator and the endowment of the anonymously-paired other subject. The subjects are not informed that their endowments are determined by choices of subjects in treatment A to avoid suggesting *indirect* reciprocity towards other subjects.

2.5 *Conclusions about Behavior*

Figure 2 shows behavior in Treatments A and B in the experiment reported in Cox (2004). This experiment was run with a double-blind payoff protocol in which the responses by individual subjects are anonymous to both the other subjects and the experimenter. Comparing the amounts sent in treatments A and B, one observes that more subjects send zero in the first mover dictator control (Treatment B) than in the investment game. Furthermore, more subjects send half ($5) or all ($10) of their endowments in the investment games than in Treatment B. So there is indeed a quite noticeable difference. Several parametric and non-parametric tests of these data support the conclusion that behavior in the investment game is known to exhibit trust because first movers send significantly more in the investment game (Treatment A) than in the first mover dictator control treatment (Treatment B). Thus behavior in the investment game is known to exhibit trust *because* first movers send significantly more in the investment game than in the first-mover dictator control treatment.

Figure 3 shows data for Treatments A and C. If one looks at the difference between the bars representing data from Treatments A and C, one sees a lot more solid bars of greater height, which suggests that play in the investment game is characterized by positive reciprocity. Several parametric and non-parametric tests support the conclusion that behavior in the investment game does exhibit positive reciprocity because second movers return significantly more in the investment game (Treatment A) than in the second mover dictator control treatment (Treatment

C). Thus, behavior in the investment game is known to exhibit positive reciprocity *because* second movers returned significantly more in the investment game than in the second mover dictator control treatment.

2.6 *Implications for Theory*

This experiment has several implications for theoretical modeling. The first is that consistency with behavior requires theory to incorporate altruistic other-regarding preferences. The reason is that the majority of subjects send positive amounts of money to another in the dictator control treatments. Furthermore, data-consistent theoretical models must incorporate beliefs about others' behavior because the triadic design reveals trusting behavior in the investment game. Finally, data-consistent models must incorporate other-regarding preferences that are conditional on the actions of others because the triadic design reveals positively reciprocal behavior in the investment game.

3. Conclusions from Other Experiments with Game Triads

The investment game is the first of several fairness games that my co-authors and I have experimented with using triadic designs. Another is the moonlighting game (Cox, Sadiraj, and Sadiraj, 2006), which is an extension of the investment game in which both first and second movers can take money as well as give it.[3] Similarly to our finding for the investment game, we conclude that behavior in the moonlighting game exhibits both positive reciprocity and trust by comparing behavior in the central game of interest with behavior in appropriately-designed dictator control treatments. Unlike the investment game, the moonlighting game can elicit negative reciprocity and fear (of negative reciprocity) because subjects can take money from each other. Data from the moonlighting game and dictator controls provide weak support for negative reciprocity and fear because some test results are significant and others are not.

The trust game is a simplified version of the investment game in which the first mover can either "exit" (which corresponds to sending zero in the investment game) or "engage," that is to say move down the game tree, in which case the second mover has choice between keeping all of an increased total payoff or sharing it in one specific way.[4] Behavior in the trust game is invariant with a doubling of money payoffs (Cox and Deck, 2005). Positive reciprocity is significant in the trust game with a single blind protocol but not with a double blind protocol (Cox and Deck, 2005). This result has possible implications for understanding behavior in other games for which experimenters have *only* used single blind protocols.

In a single blind protocol, other subjects in an experiment cannot identify what a specific individual subject has done. In a double blind protocol, neither other subjects nor the experimenter can identify what any individual subject has done. Thus if the second mover, for example, wants to defect and keep all the money, that second mover does not have to worry about being frowned upon, or worse, perhaps not invited to be in future experiments or whatever else subjects might imagine, if the experimenter uses a double blind payoff protocol. In contrast, consider the implications of a single-blind protocol in a fairness game. For illustration consider the possible case in which a first mover has sent his entire $10 endowment to the paired second mover. And suppose that the second mover considers keeping all of the $40 and leaving the paired first mover with $0. In a typical single blind protocol, the defecting second mover would be called by name to collect his $40 in a face-to-face interaction with the experimenter. Furthermore, the experimenter is typically a professor, and a professor is arguably an authority figure for student subjects. The knowledge that subjects will have to face the experimenter to collect their payoffs does dissuade some potential defectors from defecting in the trust game. Since a large majority of experiments with fairness games have been run with single blind protocols, our finding may imply that some rethinking about conclusions is needed. One cannot know, *a priori,* all of the contexts in which a double blind protocol might yield different behavior than a single blind protocol. If one observes reciprocity in a double blind experiment then it is a

really strong result which indicates that the norm for reciprocity is "internalized." In contrast, if reciprocity is observed in a single blind experiment, but not in an otherwise identical double blind experiment, then one needs to revisit the question of what the experimenter is attempting to measure because the experimenter-as-observer would have been shown to be a significant treatment. Furthermore, different experimenters may themselves have significant treatment effects: the prospect of collecting money payoffs resulting from defection from an old professor (arguably a father or mother figure) may be more constraining than the prospect of collecting such payoffs from a young graduate assistant.

Returning to results from experiments reported in our previous papers, one notes further conclusions as follows. We found that negative reciprocity and fear are not significant in the punishment mini-ultimatum game (Cox and Deck, 2005), which is a simplified version of the traditional ultimatum game.[5] We found that play in the punishment mini-ultimatum game is invariant with framing the task as market exchange (Cox and Deck, 2005). We also found that negative reciprocity is significant in the punishment mini-ultimate game *if* it is embedded within a context of similar games but not when played in isolation (Cox and Deck, 2005). This last finding is actually a little bit troubling for developing theoretical models in this area; it shows that it can indeed be quite a bit more complicated than we would like it to be. We also found that females are less positively reciprocal in investment games than are males, and that groups are less generous in the investment game than are individuals (Cox, 2002). Cox and Deck (2006) studies gender differences using a triadic experimental design including the trust game. The data indicate that women are more sensitive than men to the costs of generous actions. The factors that affect the level of observed generosity are reciprocal motivation, the level of money payoffs, and the level of social distance in the experimental protocol. The relatively greater sensitivity of women to the costs of generous behavior can explain much of the apparent inconsistencies among gender-difference experiments of previously reported in the literature. Cox and Deck (forthcoming) reports a trust game with first mover trembling, which is a game in which "nature"

randomly determines whether a first mover's decision is implemented or reversed. Data from this trust game with trembling indicate that second movers give first movers the benefit of the doubt in reacting to realization of the ungenerous branch of the game tree. However, first movers do not anticipate this forgiving response by second movers and are less likely to pursue the mutually beneficial outcome when there is trembling.

4. Models of Unconditional Other-regarding Preferences

Now I want to switch to my second theme and look at some specific models of social preferences. In models of inequality aversion, utility is increasing with one's own money payoff but decreasing with the difference between one's own and others' money payoffs.[6] In the quasi-maximin model, utility is increasing with an agent's own money payoff, with the lowest of all agents' payoffs (the maximin property), and with the total of all agents' payoffs (the efficiency property).[7] An alternative model motivated by data is the egocentric altruism model (Cox and Sadiraj, 2006), which contains other-regarding preferences that are characterized by monotonicity, convexity, and egocentricity (defined below).

4.1 *A Direct Test of Inequality Aversion*

A direct test of inequality aversion is provided by the first-mover dictator control treatment for the investment game triad with the following design (Cox and Sadiraj, 2006). The dictator is given $10. The anonymously-paired subject is given $10. The dictator can keep all of his $10 or give any integral part of it to the paired person. Any amount given is tripled by the experimenter. Behavior in this experiment was as follows. First, 19 of 30 or 63% of the dictators gave positive amounts to the other person. The average amount given was $3.63. The average payoff of dictators was $6.37 and the average payoff of non-dictators was $20.89, which implies a high degree of inequality favoring the other person. The inconsistency with the inequality aversion models does not just reflect an inconsistency with the parametric forms of these models;

instead, it is a fundamental inconsistency with inequality aversion, *per se*. The behavior of the 37% of subjects that is consistent with inequality aversion is also consistent with self-regarding (or economic man) preferences, which is the preferred model because it is the simpler of the two. Therefore, inequality aversion is not needed to rationalize the behavior of any subjects in this experiment.

4.2 Direct Tests of the Quasi-Maximin Model

Cox and Sadiraj (2006) report two direct tests of the quasi-maximin model with specially-designed dictator experiments. In each experiment, a dictator is given a choice among three rows of a table containing payoffs for herself and three other people. In one experiment, the dictator's own payoff and the minimum payoff is the same in all three rows but the total payoff varies between rows. In the other experiment, the dictator's own payoff and the total payoff is the same in all three rows but the minimum payoff varies between rows. The experiment results are that the choices of 85% of the subjects in one experiment and of 94% of the subjects in the other experiment are inconsistent with quasi-maximin preferences.

4.3 More Information about Subjects' Preferences

The dictator experiment discussed in section 4.1, that provides a direct test of inequality aversion, reveals that a high majority of subjects behave like altruists when faced with the choice between choosing zero and giving a positive amount to the paired subject when the price of each $1 given is 33 cents. But this experiment leaves open the question of how subjects behave when they can either give or take money from another. Will they still appear to be altruists?

In experiment 4 of Cox and Sadiraj (2006), a subject can choose zero or give money to the other subject or take money from him. The price of each $1 increase in the other subject's payoff is 33 cents. Each $1 taken from the other subject increases the dictator's payoff by $1. Thus the experiment reviewed here differs from the experiment reviewed in section 4.1 *only* by

introduction of the opportunity to take money as well as give it or choose zero. This is the first-mover dictator control experiment for the moonlighting game (Cox, Sadiraj, and Sadiraj, 2006). Data from this experiment are strikingly different from data for the experiment reviewed in section 4.1: the presence of the opportunity to take money causes a large majority of subjects to do just that; in fact, 69% of the subjects took money from the other person and 56% of them took the maximum possible amount of $5. Thus, in the absence of an opportunity to take money a high majority of subjects appear to be altruists but in the presence of opportunities to either give or take money a high majority of subjects appear to be selfish. Is this behavior contradictory?

4.4 The Egocentric Altruism Model

Behavior in these two dictator experiments can be rationalized by a model of other-regarding preferences with conventional properties known as the egocentric altruism model (Cox and Sadiraj, 2006). A utility function $u(m, y)$ defined on the dictator's ("my") money payoff m and the paired subject's ("your") money payoff y that is monotonically increasing in both payoffs, has indifference curves that are strictly convex to the origin, and exhibits "egocentrism" can rationalize the data. Egocentrism is defined as $u(b,a) > u(a,b)$, for all *a* and *b* such that $b > a \geq 0$; in words, the individual is assumed to be an altruist but not a "Mother Teresa." An additional regularity property can be assumed for the model and maintain consistency with data described above: the utility function is assumed to be CES, hence homothetic, which implies that slopes of indifference curves are constant along rays from the origin, hence preferences over relative income m/y are defined in a straightforward way. The egocentric altruism model is consistent with almost all of the data from all four of the dictator experiments described in this section of the paper (Cox and Sadiraj, 2006). Furthermore, this model is robust, it can also explain data from experiments with proposer competition and responder competition (Cox and

Sadiraj, 2006) and data from experiments with voluntary contributions to public goods (Cox and Sadiraj, 2005).

5. Incorporating Intentions into a Model of Social Preferences

As explained above, the egocentric altruism model can explain data from several types of dictator experiments while the inequality aversion and quasi-maximin models cannot explain such data. But neither the egocentric altruism model nor the other models incorporate intentions. Furthermore, this limitation is known to have empirical relevance because of experiments that identify the significance of reciprocity in various contexts, including experiments with the investment game (Cox, 2002, 2004), the trust game (Cox and Deck, 2005, 2006), and the moonlighting game (Cox, Sadiraj, and Sadiraj, 2006).

An implication of the parsimony objective of theoretical modeling is that intentions should be incorporated into a model that can rationalize data from experiments *without* reciprocal motivation, such as dictator games, rather than proceeding in an orthogonal direction to develop unrelated models to explain intentions-conditional behavior such as reciprocity. This approach leads to development of a unified body of theory for modeling both less and more complicated instances of revealed social preferences.

5.1 *A Parametric Model of Reciprocity and Fairness*

The egocentric altruism model is extended to incorporate intentions in the "tractable model of reciprocity and fairness" (Cox, Friedman, and Gjerstad, forthcoming) by assuming that the parameter weight that applies to another's money payoff is not an exogenous constant but, instead, is given by a function of a reciprocity variable r and a status variable s that are dependent on the other person's revealed intentions and social status characteristics that are relevant to the decision environment. The resulting utility function is a modified CES function of

a decision-maker's own ("my") money payoff m and the other's ("your") money payoff y with a multiplicative weight for y given by the weighting function $\theta(r,s)$. The marginal willingness to pay to increase the other's payoff, when it is equal to one's own, is equal to $\theta(r,s)$. This θ function is assumed to be weakly increasing in both arguments, to have the neutral-state property given by $\theta(0,0) \geq 0$, and to be negative when r and s are sufficiently negative. Thus individuals are assumed to non-malevolent in their baseline state of $(r,s) = (0,0)$, to be benevolent if the reciprocity and status variables are sufficiently positive, and to be malevolent if the reciprocity and status variables are sufficiently negative. In applying the model to data, the θ function is assumed to be identical across individuals except for a mean zero idiosyncratic term; in other words, individual agents are allowed to differ in their baseline altruism.

Data used in estimating the model come from several distinct types of experiments. Application of the model to data from the baseline dictator game, with random role assignment, reported by Cherry, Frykblom, and Shogren (2002) yields estimates of individuals' residual or baseline altruism. Effects of experimenter-induced status on altruism are derived by applying the model to data from the dictator game with earned endowments reported by Cherry, Frykblom, and Shogren (2002). Estimates of subjects' reciprocity are derived by applying the model to data from Stackelberg duopoly (Huck, Muller, and Norman, 2001) and moonlighting games (Cox, Sadiraj, and Sadiraj, 2006). Reciprocity with context-dependent property rights is studied by applying the model to mini-ultimatum game data reported by Falk, Fehr, and Fischbacher (2003). Effects of reciprocity and status on subjects' other-regarding behavior are derived by applying the model to data from ultimatum games with random and contest assignment of the first-mover role, reported by Hoffman, McCabe, Shachat, and Smith (1994). Individuals' efficiency-increasing behavior is studied by applying the model to data from the first-mover dictator control treatment in the investment game triadic-design experiment reported in Cox (2004) and Cox and Sadiraj (2006).

5.2 A Non-parametric Model of Revealed Altruism

The egocentric altruism model and tractable model of reciprocity and fairness are further generalized in a non-parametric model based on partial orderings of preferences and opportunity sets (Cox, Freidman, and Sadiraj, 2006). In this model, one agent's other-regarding preferences can depend on the actions of the other agent. The model is based on two partial orderings and two axioms that link them. The partial ordering of preferences is a formal representation of what it means for one preference ordering to be "more altruistic than" another. The partial ordering of opportunity sets is a formal representation of what it means for one opportunity set to be "more generous than" another. These two partial orderings are linked by two axioms. Axiom R states that more generous choices by the first mover in an extensive form game induce more altruistic preferences in the second mover. Axiom S states that the reciprocity effect on preferences is stronger following an act of commission by the first mover than following an act of omission. This non-parametric model is applied to data from the investment game, Stackelberg duopoloy game, and Stackelberg mini-game.

6. Summary

This research program involves experiments with "fairness games" designed to reveal the characteristics of individuals' social preferences and an approach to modeling these social preferences based on application of the objective of parsimony. The experiments reveal that behavior in fairness games exhibits unconditional altruism ("others' payoffs matter"), trust ("beliefs matter"), and reciprocity ("intentions matter"). Whether or not reciprocity is exhibited in some games depends upon whether the experimenter uses a single-blind or double-blind protocol ("who is observing matters") and the context in which a specific game is embedded ("fairness is a relative concept"). The experiments reveal that other-regarding behavior differs

across small groups and individuals and across males and females (“type of decision-maker matters”).

Modeling behavior in fairness games involves complications that vary with characteristics of the games. In simple dictator games that do *not* elicit reciprocal motivation, behavior is *inconsistent* with inequality aversion and quasi-maximin preferences. Behavior in these dictator games and in games of proposer competition, responder competition, and voluntary contributions to public goods can be rationalized by a model of egocentric altruism. Behavior in games such as the investment, trust, moonlighting, ultimatum, mini-ultimatum, Stackelberg duopoly, and Stackelberg mini-games that *do* elicit reciprocal motivation can be modeled with a tractable parametric extension of the egocentric altruism model and with a non-parametric revealed altruism model based on a partial ordering of preferences (“more altruistic than”), a partial ordering of opportunity sets (“more generous than”), and “reciprocity” and “status” axioms that link the two partial orderings.

Footnotes

* Noah Langdale Jr Eminent Scholar Chair and Director of the Experimental Economics Center (EXCEN), Georgia State University. This paper was written while the author was a Visiting Scholar at the Workshop in Political Theory and Policy Analysis, Indiana University.

1. Fehr and Gächter (2000) survey some of this literature.

2. Models of inequality averse preferences are presented by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000).

3. The moonlighting game was introduced to the literature by Abbink, et al. (2000).

4. The trust game was introduced to the literature by McCabe and Smith (2000).

5. Mini-ultimatum games were previously experimented with by Bolton and Zwick (1995), Gale, et al. (1995), and Falk, et al. (2003).

6. See Fehr and Schmidt (1999) and Bolton and Ockenfels (2000).

7. The quasi-maximin model was introduced to the literature by Charness and Rabin (2003).

References

Abbink, Klaus, Irlenbusch, Bernd, and Renner, Elke (2000), "The Moonlighting Game: An Empirical Study on Reciprocity and Retribution." *Journal of Economic Behavior and Organization*, 42, pp. 265-77.

Bolton, Gary E. and Ockenfels, Axel (2000), "ERC: A Theory of Equity, Reciprocity and Competition." *American Economic Review*, 90, pp. 166-93.

Bolton, Gary E. and Zwick, Rami (1995), "Anonymity versus Punishment in Ultimatum Bargaining," *Games and Economic Behavior*, 10, pp. 95-121.

Charness, Gary and Rabin, Matthew (2002), "Social Preferences: Some Simple Tests and a New Model," *Quarterly Journal of Economics*, 117, pp. 817-69.

Cox, James C. (2002), "Trust, Reciprocity, and Other-Regarding Preferences: Groups vs. Individuals and Males vs. Females," in R. Zwick and A. Rapoport, (eds.), *Experimental Business Research*, vol. I, Kluwer Academic Publishers.

Cox, James C. (2003), "Trust and Reciprocity: Implications of Game Triads and Social Contexts," University of Arizona Discussion Paper Number 00-11.

Cox, James C. (2004), "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 46, no. 2, pp. 260-281

Cox, James C. and Deck, Cary A. (2005), "On the Nature of Reciprocal Motives," *Economic Inquiry*, 43, no. 3, pp. 623 – 635.

Cox, James C. and Deck, Cary A. (2006), "When are Women More Generous than Men?", *Economic Inquiry*, in press.

Cox, James C. and Deck, Cary A. (forthcoming), "Assigning Intentions when Actions are Unobservable: the Impact of Trembling in the Trust Game," *Southern Economic Journal*.

Cox, James C., Friedman, Daniel and Gjerstad, Steven (forthcoming), "A Tractable Model of Reciprocity and Fairness," *Games and Economic Behavior*.

Cox, James C., Friedman, Daniel, and Sadiraj, Vjollca (2006), "Reciprocal Altruism," Georgia State University working paper.

Cox, James C., Sadiraj, Klarita, and Sadiraj, Vjollca (2006), "Implications of Trust, Fear, and Reciprocity for Modeling Economic Behavior," University of Arizona Discussion Paper 6, 2001, revised 2006.

Cox, James C. and Sadiraj, Vjollca (2005), "Social Preferences and Voluntary Contributions to Public Goods," paper presented at the Experimental Public Economics Conference, Andrew Young School of Policy Studies, Georgia State University.

Cox, James C. and Sadiraj, Vjollca (2006), "Direct Tests of Models of Social Preferences and a New Model," Georgia State University working paper.

Falk, Armin., Fehr, Ernst, and Fischbacher, Urs (2003), "On the Nature of Fair Behavior." *Economic Inquiry*, 41, pp. 20-6.

Fehr, Ernst and Gächter, Simon (2000), "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, Summer, 14, pp. 159-81.

Fehr, Ernst and Schmidt, Klaus M. (1999), "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, pp. 817-68.

Gale, John, Binmore, Kenneth G. and Samuelson, Larry (1995), "Learning to Be Imperfect: The Ultimatum Game," *Games and Economic Behavior*, 8, pp. 56-90.

McCabe, Kevin A., and Smith, Vernon L. (2000), "A Comparison of Naïve and Sophisticated Subject Behavior with Game Theoretic Predictions." *Proceedings of the National Academy of Sciences*, 97, pp. 3777-81.

Samuelson, Paul A. (1947), *Foundations of Economic Analysis* (Harvard University Press, Cambridge, MA).

Figure 1 : Amounts Sent and Returned in Treatment A

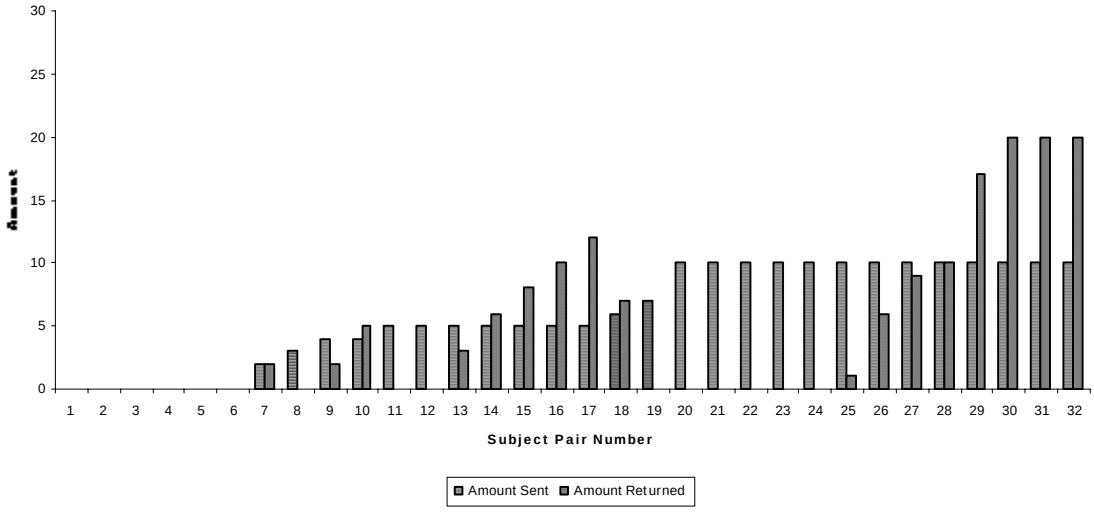

Figure 2 : Amounts Sent in Treatments A and B

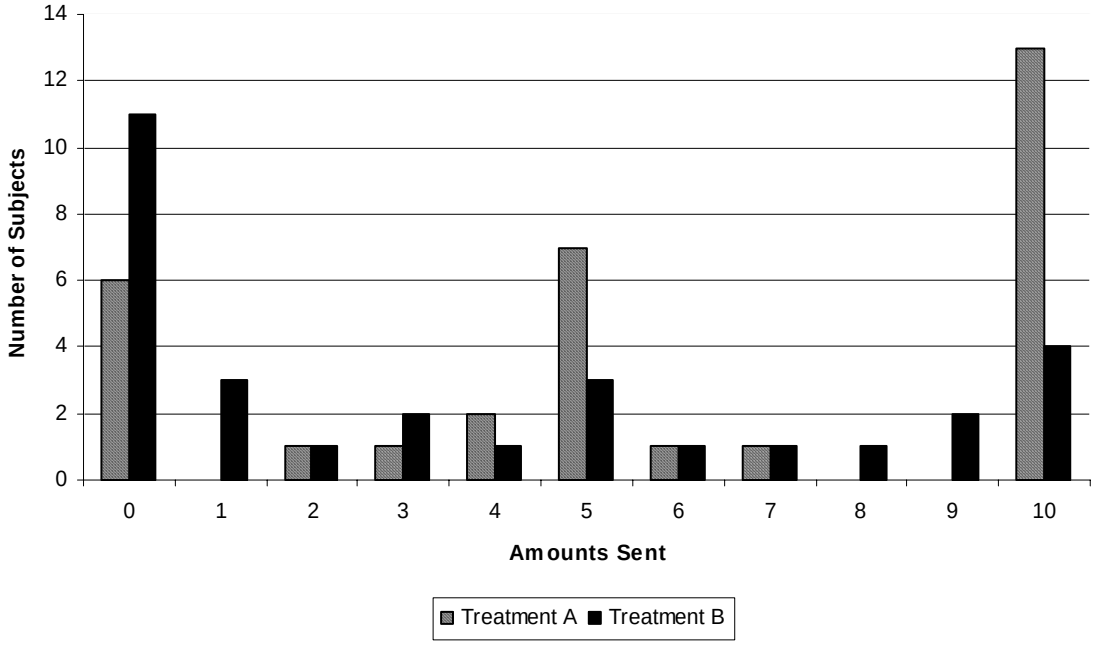

Figure 3 : Amounts Returned in Treatments A and C