Trust, Fear, Reciprocity, and Altruism

James C. Cox
University of Arizona
jcox@bpa.arizona.edu

Klarita Sadiraj
Nyenrode Forum for Economic Research
k.sadiraj@nyfer.nl

Vjollca Sadiraj
University of Amsterdam
vjollca@fee.uva.nl

August 2001; revised October 2002

Trust, Fear, Reciprocity, and Altruism*

This paper uses a triadic experimental design to discriminate between actions motivated by (intentions-unconditional) preferences over the distribution of material outcomes and actions motivated by attributions of the intentions of others. Such discrimination is essential to empirical guidance for theory development because modeling intentions is quite different than modeling (intentions-unconditional) preferences. The triadic design includes the moonlighting game in which first-mover actions can motivate positively- or negatively-reciprocal actions by second movers. First movers can be motivated by trust or fear, in addition to selfish or altruistic preferences. Second movers can be motivated by altruistic, inequality-averse, or selfish preferences as well as positive or negative reciprocity. The triadic design includes specially-designed dictator control treatments to discriminate among actions with alternative motivations. Data from the experiment support the conclusion that first movers' behavior in the moonlighting game is characterized by trust and an absence of fear. Furthermore, the first movers' behavior is based on realistic expectations because the second movers' behavior is characterized by positive reciprocity but not by significant negative reciprocity.

Keywords: game theory, trust, fear, reciprocity, altruism
JEL Classification: C70, C91, D63, D64

1. Introduction

Applications of game theory have historically focused on the model of "self-regarding preferences" in which agents are assumed to be exclusively concerned with maximizing their own material payoffs. This model predicts behavior quite well in many types of controlled experiments. But there is now a large body of experimental literature that has produced replicable patterns of inconsistencies with the self-regarding preferences model's predictions in contexts involving salient fairness considerations or opportunities for cooperation. This literature was reviewed in a recent survey paper on "the economics of reciprocity" by Fehr and Gächter (2000).

Actions that are inconsistent with the predictions of the self-regarding preferences model can be motivated by social norms for reciprocating the intentional actions of another. But actions that are inconsistent with self-regarding preferences can also be motivated by agents' altruistic or inequality-averse (intentions-unconditional) preferences over the distribution of material payoffs.[1]

The distinction between actions motivated by intentions-unconditional preferences over outcomes and actions motivated by attributions of intentions is essential to empirical guidance for theory development because modeling intentions is quite different than modeling preferences over outcomes that are unconditional on perceived intentions.[2]

Our research question is to ask whether attribution of intentions is a significant motive for behavior in experimental two-person extensive-form games, not whether such attribution is or is not a characteristic of human behavior in all contexts. But everyday life provides much anecdotal evidence that attribution of intentions, as well as preferences over outcomes, is important in social, political and economic exchange. Thus a spouse, date, or guest who is late to dinner or some other engagement is more likely to be easily forgiven if he can make a credible case that his tardiness was caused by events that were largely outside his control. A price increase by a seller is more likely to be accepted without grumbling or retaliation by buyers if the seller can credibly claim that the price increase was "necessitated" by an increase in costs rather than chosen to increase profit after an increase in demand or decrease in competitors' supply. A politician who adopts a policy that is harmful to the perceived self-interest of some constituents is more likely to survive in office if she can credibly claim that the decision was necessitated by international treaty, the political opposition, or fiscal realities. Attribution of intentions is important in both criminal and civil law. Thus the distinction in law between the crimes of manslaughter and murder turns on the intent of the perpetrator. In the civil law, whether or not punitive damages are awarded, and if so their amount, depends on the perceived intentions of the defendant.

In order to obtain data that can guide development of economic and game-theoretic models, we need to be able to discriminate between actions with alternative motivations. We use a three-games or "triadic" experimental design to discriminate between actions motivated by intentions-unconditional preferences over outcomes and actions motivated by attributions of intentions in the moonlighting game. The moonlighting game was introduced to the literature by Abbink, Irlenbusch, and Renner (2000); it is an extension of the investment game of Berg,

Dickhaut, and McCabe (1995). The triadic design uses specially-constructed dictator games, as control treatments, that eliminate attribution of the other subject's intentions and expectations about the other subject's behavior because the paired subject has no choice to make.

Results from earlier experiments with the moonlighting game by Abbink, Irlenbusch, and Renner (2000) are inconsistent with the self-regarding preferences model and consistent with behavior motivated by attributions of intentions such as positive and negative reciprocity. But their data are also consistent with behavior motivated by altruistic and inequality-averse preferences over outcomes that are not conditional on attributions of others' intentions. A more elaborate experimental design is needed to discriminate between alternative motivations for behavior in the moonlighting game.

There are a few previous studies that used control treatments for intentions. Blount (1995) reported results from an experimental design in which second-mover rejections in a standard ultimatum game were compared with second-mover rejections in games in which the first move was selected randomly or by an outside party rather than by the subject that would receive the first mover's monetary payoff. She found lower rejection rates in the random treatment than in the standard ultimatum game but no significant difference between the rejection rates in the third-party and standard games. Charness (1996) introduced Blount's control treatments into a design for experiments with the gift exchange game. He reported that second mover contributions in the standard gift exchange game were somewhat *lower* than in the outside party and random treatments. Bolton, Brandts, and Ockenfels (1998) used an intentions-control treatment in their design for experimenting with simple dilemma games. In the control treatment, the row player is given the task of "choosing" between two identical rows of monetary payoffs. They found no significant differences between the column players' responses in the control treatments and the "positive and negative reciprocity" treatments. Cox (2002a, 2002b) used a triadic experimental design for experiments with the investment game. One dictator game controlled for the effects of the second mover's possible attribution of the first mover's intentions. The other dictator game

controlled for the effects of the first mover's possible expectations about the second mover's decisions. The data from the triadic-design experiments with the investment game support the conclusions that first-mover behavior is motivated by trust, as well as altruism, and second-mover behavior is motivated by positive reciprocity, as well as altruism.

Our triadic experimental design, incorporating the moonlighting game, can discriminate among first-mover choices motivated by trust, fear, or altruism. This design can also discriminate between second-mover decisions motivated by positive reciprocity or altruism and between second-mover choices motivated by negative reciprocity or inequality aversion.

2. Experimental Design and Procedures

The experiment sessions were run with custom computer software in the CREED laboratory at the University of Amsterdam in the fall of 2000. The experiment included three treatments implemented in an across-subjects design. All money payoffs and subjects' feasible choices were quoted in numbers of euro.[3] At the time the experiment sessions were run, 1 euro was worth a little less than 1 dollar. At that time, the euro was not yet a circulating currency but prices in retail stores were quoted in both Guilders and euro. The subjects were paid in Guilders, using the official exchange rate of 2.20 Guilders per euro. The subject instructions included the exchange rate, which would in any case have been known by the subjects from retail shopping experience. The payoffs and feasible choices were quoted in numbers of euro, rather than Guilders, in order to make subjects' economic incentives about the same as in earlier investment game experiments while, at the same time, making their endowments of 10 currency units and unit of divisibility of one currency unit comparable to the $10 endowments and $1 unit of divisibility used in those earlier experiments.[4]

2.1 The Three Games

Treatment A is the moonlighting game. Each individual in the second-mover group is credited with a 10 euro endowment. Each individual in the first-mover group is credited with a 10 euro endowment and given the task of deciding whether she wants to give to a paired individual in the other group none, some, or all of her 10 euro or take up to 5 euro from the paired person. Any amounts given by the first mover are tripled by the experimenter. Any amounts taken by the first mover are not transformed by the experimenter. Then each individual in the second-mover group is given the task of deciding whether he wants to give money to the paired first mover or take money from her. Each euro that the second mover gives to the paired first mover costs the second mover 1 euro. Each euro that the second mover takes from the paired first mover costs the second mover 1/3 euro. The second mover's choices are constrained so as not to give either mover a negative payoff. All choices by first movers and second movers in all treatments are required to be in integer amounts.

Figure 1 shows representative choices open to the subjects in treatment A. The piece-wise-linear solid line passing through points A , I, and D is the first mover's "budget line." The two subjects' endowments are at point I, the intersection of the first mover's "budget line" and the 45-degree line. The slope of the first mover's "budget line" is -3 above the 45-degree line and -1 below the 45-degree line. The first mover's choice of an integer amount to give to or take from the second mover determines the second mover's "budget line." If the first mover were to give 7 euro to the second mover, then the second mover's "budget line" would be the piece-wise-linear dashed line with a kink at point A in Figure 1. The slope of the second mover's "budget line" in Figure 1 is $+1/3$ below the first mover's budget line and -1 above it. If the first mover were to take 3 euro from the second mover then the second mover's "budget line" would be the piece-wise-linear dashed line with a kink at point D in Figure 1.

Thus, in treatment A, the first mover chooses a "budget line" for the second mover by choosing an integer amount to give to or take from the second mover. Subsequently, the second mover chooses a point on this "budget line" that determines both first- and second-mover money payoffs. The second mover's choice is an integer amount to give to or take from the first mover.

Treatment B is a dictator game that differs from treatment A only in that the individuals in the "second-mover" group do not have a decision to make. Thus, in treatment B the first mover chooses a point on the first-mover "budget line" in Figure 1. This choice determines the money payoffs for both subjects.

Treatment C is a dictator game that involves a decision task that differs from treatment A as follows. First, the "first movers" do not have a decision to make. Each "second mover" is given one of the "budget lines" determined by a first mover's decision in treatment A. The "second movers" are *not* informed about the source of the "budget lines." The "second mover" then determines both subjects' money payoffs by choosing a point on her "budget line."

2.2 Procedures

The subjects assembled in a sign-in room. They registered on a subject list and picked up copies of printed instructions from a stack on a table. The subjects drew small envelopes and folded "notes" from two different boxes, each containing items that were identical on the outside. Each envelope contained a mailbox key with a unique identification code. The subjects were asked not to open their envelopes until they were seated at computers in the laboratory. The key codes were to be used for subject identification for money payoffs. One-half of the notes contained the symbol # and one-half contained the symbol $*$. The random assignment of symbols on the notes implemented the random assignment of subjects to the two sections of the laboratory.

Subsequently, the subjects walked a few feet down the hallway and entered the laboratory through either the door marked with # or the door marked with $*$. The experimenters stood in the hallway, well back from the two doors, and in a position where observation of which subject

approached which computer was impossible. After all subjects had entered the laboratory, the doors were closed for the duration of the experiment. The laboratory was divided into two sections by a floor to ceiling partition. One section was accessed through the door marked # and the other section was accessed through the door marked ∗. The windows between the experimenters' control room and the laboratory were covered by blinds. Thus, the two groups of subjects had no verbal, visual, or other contact with each other or with the experimenters during the decision-making part of the experiment.

The subjects read the instructions on their computer monitors. A printed version of the instructions was also available on each subject's computer monitor desk in case the subject wanted to review the instructions during the decision-making part of the experiment. The instructions referred to the subjects only as being in group X or group Y. Terms such as first mover, second mover, proposer, responder, etc. were avoided. The instructions stated that subjects could "increase" or "decrease" their own and the paired subject's "account balances." The instructions did not use the words "send" and "return" for the amounts transferred by first and second movers. Other, possibly more evocative verbs, such as "give," "take," "reward," and "punish" were avoided. Tables in the instructions presented all feasible actions and their consequences for both subjects in a pair of first and second movers.

The end of the on-screen instructions directed the subjects to enter their key codes into their computers and then proceed to answer the questions that would appear on their computer monitors. The questions were intended to test subjects' understanding of the experimental tasks and procedures. If a subject answered a question incorrectly, she was informed of this by a message on her computer screen that also asked her to try again. After all of the subjects answered all questions correctly, the decision-making part of the experiment began. An English translation of the Dutch instructions given to the subjects is available on an author's webpage.[5]

The decision-making part of the experiment proceeded as follows. First, the monitor computer randomly determined which room, # or ∗ was the room with group X subjects and

which was the room with group Y subjects. The pairing of group X and group Y subjects was established by where the subjects sat in the two separated parts of the laboratory. Thus the subjects had no way of knowing who they were paired with. And the experimenters had no way of knowing which subject sat at which computer. Salient payoffs were possible because the subjects entered their key codes in their computers. The payoff procedure was double blind: (a) subject responses were identified only by the key codes that were private information of the subjects; and (b) money payoffs were collected in private from sealed envelopes contained in coded mailboxes.

The decision task in treatment C was implemented as follows. Each subject pair, j = 1,2,…,30, was informed that the person in group X had a beginning account balance of $10 + A_j$ and the person in group Y had a beginning account balance of $10 + B_j$. The amounts, A_j and B_j were determined by first movers' decisions in treatment A. A subject pair in treatment C was informed of the amounts, A_j and B_j but not told that they had been determined by the decision of the group X person in subject pair j in treatment A. The decision to withhold this information was based on the judgment that it might motivate indirect reciprocity by the subjects, which would be inappropriate in this control treatment.[6] A desire to avoid an alternative type of indirect reciprocity also accounts for the way the endowments were implemented. A different procedure than we used would be to first endow each subject in every pair with 10 euro and, subsequently, have the experimenter or another third party alter the endowments for pair j by A_j and B_j. This alternative procedure would involve "level 2 attribution," with perceptions of intentionality but not self-interest (Blount, 1995, p.113). Our treatment C procedure is "level 3 attribution," which removes perceptions of both intentionality and self-interest. This provides the comparison we want with treatment A, which is "level 1 attribution" involving perceptions of both intentionality and self-interest. Thus, comparison of data from treatment A with data from

treatment C provides a measure of the incremental effect of direct reciprocity on subjects' decisions that is not confounded by the possible effect of indirect reciprocity.

All of the above design features were common information given to the subjects except for the aforementioned withholding of the source of the A_j and B_j figures in treatment C. Each treatment was run in four sessions. There were never fewer than 12 nor more than 18 subjects in a session. The experimental treatments were implemented "across-subjects"; that is, different subjects participated in each of the three treatments.

3. Discriminating Among Trust, Fear, Reciprocity, and Other-Regarding Preferences

As with any data, one needs a maintained theoretical model to interpret the data from the moonlighting game triadic experiment. The triadic design provides data that can be used to discriminate empirically among choices motivated by unconditional preferences over outcomes and choices motivated by attributions of another's previously-revealed intentions or beliefs about another's future actions. We consider a general model and use the data to inform us about which properties of the model receive empirical support.

Assume that each subject in every pair has preferences over her own and the paired subject's money payoffs that can be represented by a utility function. These preferences can be self-regarding or other-regarding. If the preferences are self-regarding then the utility function is a constant function of the other's money payoff and strictly increasing in one's own money payoff. If the preferences are other-regarding then they can be inequality-averse (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) or altruistic (Cox, Sadiraj, and Sadiraj, 2002). The preferences are inequality averse if the utility function is decreasing in the other person's money payoff when his payoff is higher than one's own. The preferences are altruistic if the utility function is globally increasing in both one's own money payoff and the other's money payoff.

The altruistic preferences are "egocentric" if in choosing between payoff pairs, $(m, y) = (a, b)$ and $(m, y) = (b, a)$, where m is "my payoff" and y is "your payoff," the agent prefers (a, b) to (b, a) whenever $a > b$ (Cox, Sadiraj, and Sadiraj, 2002). The preferences may include expectations about the future actions of another. The preferences over outcomes may be conditional on attributions of the intentions of another that are revealed by the other's preceding action.

3.1. Identifying Altruistic and Trusting Behavior

A first mover will be said to undertake an action that exhibits trust if the chosen action: (a) sends the second mover a positive amount of money, which creates a monetary gain that could be shared; and (b) exposes the first mover to the risk of a loss of utility. Thus a trusting action requires a belief by the first mover that the second mover will not defect and keep too much of the profit generated by the first mover's decision to send a positive amount.

If a first mover has self-regarding preferences then she will have indifference curves in Figure 1 that are vertical straight lines. In that case, the act of sending any positive amount implies trust because such a first mover will have lower utility that at the endowment point, *I* unless the second mover returns at least as much money as the first mover gave up. If a first mover has inequality-averse preferences then he will have indifference curves in Figure 1 that have positive slopes above the 45-degree line. In that case, the act of sending any positive amount implies trust because such a first mover will have lower utility that at the endowment point, *I* unless the second mover returns at least as much money as the first mover gave up. But a first mover may have altruistic other-regarding preferences. Since, in the moonlighting game any amount sent by the first mover is tripled, a first mover with egocentric altruistic preferences might prefer to give the second mover some money regardless of how much, if any, the second mover might return (assuming that the amount returned is no larger than four times the amount sent,

which is of course true for all observations in our data). Thus the mere act of sending a positive amount of money is not evidence of trusting behavior unless it is known that first movers have self-regarding or inequality-averse preferences. But the treatment B dictator game, together with the treatment A moonlighting game, permit one to identify trusting actions by altruists, as follows.

In treatment B, a first mover chooses an amount to send from the set, S of integers weakly between -5 and 10. The choice of $s_b > 0$ implies

(1) $u^1(10 - s_b, 10 + 3s_b) \geq u^1(10 - s, 10 + 3s)$, for all $s \in S$.

Now assume that the amount of money that the first mover sends to the second mover in treatment A, s_a is both positive and larger than the amount "sent" in treatment B (which may be positive, zero, or negative). Then we can conclude that the first mover has exhibited trust because the amount sent in treatment A is too large to be fully explained by egocentric altruistic preferences. Thus, if $s_a > 0$ and $s_a > s_b$ then we know that the first mover is exposed to risk from the possibility that the second mover will defect and appropriate too much of the money transfer. If the second mover were to return nothing in the event that $s_a > s_b$, then statement (1) and strict quasi-concavity of u^1 imply that the first mover will have lower utility than he could have attained if he had known that the second mover would return nothing:

(2) $u^1(10 - s_a, 10 + 3s_a) < u^1(10 - s_b, 10 + 3s_b)$

because $s_a \in S$. Thus the sufficient condition for concluding that a first mover has exhibited trust is $s_a > 0$ and $s_a > s_b$. (See the appendix for a formal derivation.)

Figure 2 illustrates a trusting action by a first mover with altruistic other-regarding preferences. An example of an altruistic choice, $s_b > 0$ that satisfies (1) with a strict inequality is shown by the tangency of the first mover's indifference curve with his "budget line" at the point

B in Figure 2. An example of a trusting choice, $s_a > s_b > 0$ that satisfies (2) is shown in Figure 2 as point A. This choice gives the second mover the piece-wise-linear dashed budget line with a kink at point A. If the second mover chooses any amount to return that is below the first mover's indifference curve passing through point B, then the first mover is worse off than she would have been if she had, instead chosen $s_a = s_b$ (at point B in Figure 2) and the second mover had returned any amount of money from 0 to $4s_b$.

3.2. Identifying Positively-Reciprocal Behavior

Next consider the question of identifying positively-reciprocal behavior. The preferences over payoff (ordered) pairs can be conditional on a social norm for reciprocating the intentionally generous behavior of another. For example, if the second mover knows that the first mover in the moonlighting game intentionally sends the second mover some of the first mover's money, the second mover may be motivated by a social norm for reciprocity to repay this generous action with a generous response. Within the context of a model of preferences over material payoffs, a social norm for reciprocity can be introduced with a state variable. Thus, the preferences over payoffs can be conditional on a state variable for reciprocity. This is an appropriate representation because, *if* there is reciprocal behavior, *then* individuals behave as if they are more altruistic towards another person after that person has been kind, generous, or trusting. The empirical question is whether or not second movers in the moonlighting game choose more generous actions, after the first mover has intentionally sent them money, than they would in the absence of the first mover's action but the presence of the same money allocation.

When analyzing data from this experiment, we will use the following specific criterion for deciding whether a subject's behavior is positively reciprocal. A second mover will be said to undertake an action that exhibits positive reciprocity if the chosen action: (a) follows a positive transfer of money by the first mover; (b) gives the first mover a monetary gain; and (c) is

undertaken instead of an available alternative action that would produce outcomes preferred by the second mover in the absence of the action by the first mover.

A second mover with self-regarding preferences will not return any money to the first mover. But a second mover with either altruistic or inequality-averse other-regarding preferences may return money to the first mover who, after making a positive transfer to the second mover, now has a lower money endowment than the second mover. Thus the mere fact that the second mover returns money to the first mover is not evidence of positive reciprocity. But the treatment C dictator game, together with the treatment A moonlighting game, permit one to identify positively-reciprocal actions, as follows.

A "second mover" in treatment C is given an endowment that is inversely related to the endowment of the paired subject. The endowments of a pair of subjects in treatment C are determined by a (distinct) first mover's decision in treatment A (but the subjects do not know this). For $s_a > 0$, the endowments of a pair of treatment C subjects are given by $(10 - s_a, 10 + 3s_a)$. In treatment C, a "second mover" chooses an amount to return, r_c from the set,

(3) $$R(s_a) = \{m(s_a), m(s_a) + 1, \cdots, M(s_a)\}.$$

If $r_c < 0$ then the first mover's money payoff is decreased by $3|r_c|$ and the second mover's money payoff is decreased by $|r_c|$. If $r_c > 0$ then the first mover's money payoff is increased by r_c and the second mover's money payoff is decreased by r_c. The second mover is not allowed to choose a return amount that would leave the first mover with a negative monetary payoff; therefore, $r_c \geq -((10 - s^a)/3)$. Also, r_c must be an integer; therefore,

(4) $$m(s_a) = -\max(y \in Z \ni y \leq ((10 - s_a)/3)),$$

where Z is the set of integers. The second mover is not allowed to choose a return amount that would leave herself with a negative payoff; therefore $r_c \leq M(s^a)$, where

(5) $$M(s_a) = 10 + s_a, \text{ if } s_a < 0$$
$$= 10 + 3s_a, \text{ otherwise.}$$

Define the indicator variable: $I_{-r} = 1$, if $-r > 0$ and $I_{-r} = 0$, otherwise. Given $s_a > 0$, the choice of r_c in treatment C implies

(6) $$\begin{aligned} &u^2(10 + 3s_a - |r_c|, 10 - s_a + (1 + 2I_{-c})r_c) \\ &\geq u^2(10 + 3s_a - |r|, 10 - s_a + (1 + 2I_{-c})r) \end{aligned}$$

for all $r \in R(s_a)$. An example of a choice, r_c that satisfies (6) with a strict inequality is shown by the tangency of the "second mover's" indifference curve with her "budget line" at point C in Figure 2.

Suppose that the second mover returns to the first mover in the moonlighting game, treatment A, a positive amount of money or, perhaps, even a larger amount than the first mover sent, $r_a \geq s_a$. This, in itself, does not support a conclusion that the second mover was motivated by positive reciprocity because the assumed choice could have been motivated by unconditional altruistic preferences (as shown by point C in Figure 2) or inequality-averse other-regarding preferences. However, if one observes that $r_a > r_c$ then we can conclude that the second mover was motivated by reciprocity because the amount of money returned is too large to be fully accounted for by unconditional other-regarding preferences. This follows from noting that $r_a > r_c$, statement (6), and strict quasi-concavity of u^2 imply

(7) $$\begin{aligned} &u^2(10 + 3s_a - |r_a|, 10 - s_a + (1 + 2I_{-r})r_a) \\ &< u^2(10 + 3s_a - |r_c|, 10 - s_a + (1 + 2I_{-r})r_c) \end{aligned}$$

because $r_a \in R(s_a)$. In Figure 2, any location of r_a on the second mover's "budget line" that is downwards and to the right of point C would exhibit positive reciprocity. (See the appendix for a formal demonstration that $r_a > 0$ and $r_a > r_b$ is the necessary and sufficient condition for positive reciprocity.)

It might, at first, seem inconsistent with utility maximization for a subject to return an amount of money, r_a that satisfies inequality (7). But a social norm for reciprocity can change an agent's preferences over material payoffs. Such a norm can be incorporated into a theory of utility by introducing the possibility that an agent's preferences over outcomes can depend on the observed behavior of another. Specifically, with respect to reciprocity, an agent's preferences over his own and another person's material payoffs can depend on whether the other person intentionally helped him or intentionally hurt him or did neither. Thus, let λ_a be a state variable that depends on the amount of money sent by the first mover to the second mover in treatment A:

(8) $\lambda_a = f(s_a)$.

The utility to the second mover of the monetary payoffs in the moonlighting game can be conditional on the reciprocity state variable. Thus there need be no inconsistency between inequality (7) and the intentions-conditional utility function inequality,

(9) $u^2_{\lambda_a}(10 + 3s_a - |r_a|, 10 - s_a + (1 + 2I_{-r})r_a) \geq u^2_{\lambda_a}(10 + 3s_a - |r|, 10 - s_a + (1 + 2I_{-r})r)$,

for all $r \in R(s_a)$. Furthermore, experiments on reciprocal behavior can be characterized as research on the comparative properties of intentions-unconditional (u^2), and intentions-conditional ($u^2_{\lambda_2}$) utility-maximizing behavior.

3.3. Identifying Fearful and Negatively-Reciprocal Behavior

We next turn our attention to fear of punishment and negatively-reciprocal (or punishing) behavior. We will use the following specific criteria for deciding whether a first mover's behavior is fearful in the context of the three games. A first mover will be said to undertake an action that exhibits fear if, in two otherwise-identical environments, he: (a) takes money from the second mover when the second mover does not have an opportunity to retaliate; and (b) takes less money or none from the second mover when the second mover does have an opportunity to retaliate. Figure 3 illustrates behavior of a fearful first mover. The first mover's choice in treatment B is at the tangency point B in Figure 3. Assume that in the moonlighting game the first mover takes less money or none; Figure 3 shows the outcome in which the first mover chooses the endowment point, I in treatment A. Point I is on a lower indifference curve for the first mover than is point B. The choice of I in the moonlighting game is evidence that the first mover is afraid to choose his most preferred outcome point B to define the second mover's "budget line" in the moonlighting game. (See the appendix for a formal derivation.)

Figure 3 also shows an outcome for which the second mover in the moonlighting game exhibits neither positive nor negative reciprocity. Suppose that, after the first mover sends zero (chooses point I), the second mover returns zero in both treatment A and treatment C. Figure 3 shows the case where the second mover's highest attainable indifference curve for her intentions-unconditional preferences touches the upper right corner of her piece-wise-linear "budget line" at point I. If the second mover also chooses I in treatment A, her behavior reveals neither positive nor negative reciprocity.

Figure 4 illustrates a case with absence of fear and the presence of negative reciprocity. Here, the first mover is assumed to take the same amount of money in the moonlighting game and the treatment B dictator game; the choice is at point B in Figure 4. The second mover's choice in the treatment C dictator game is also assumed to be at point B, indicating an absence *of evidence* of inequality aversion. But the second mover's choice in the moonlighting game is assumed to be

between points N and B, on the upward-sloping segment of her "budget line." Such a choice means that the second mover is paying money in order to take money from the first mover; that is, this is negatively-reciprocal behavior (for the illustrated case). Of course, for the second mover's choice to be on the upward-sloping segment of her "budget line," her indifference curve must have positive slope. In other words, the second mover's intentions-conditional utility function must be decreasing in the first mover's money payoff after the first mover has chosen the hostile action of taking money from the second mover. Thus an action undertaken by a second mover exhibits negative reciprocity if the chosen action: (a) follows a non-positive transfer of money by the first mover; (b) reduces the first mover's money payoff; and (c) is an action that would, in the absence of the hurtful action by the first mover, be dispreferred by the second mover to an available alternative action. (See the appendix for a formal derivation.)

4. Subjects' Behavior in the Experiment

The subjects' behavior in treatment A, the moonlighting game is presented in Figure 5. The reported figures include the multiplication by three for positive amounts sent by first movers and for negative amounts taken by second movers. We observe that 12 out of the 30 first movers took the maximum of five euro and one subject took one euro from the paired second mover. Three subjects "sent" zero and 14 subjects gave positive amounts to the second mover; five subjects gave the maximum of 30 euro. On average, it was profitable (in monetary terms) for the first movers to give money to second movers. First movers who sent positive amounts of money to second movers made an average profit of 1.93 euro after the second movers made their return decisions. In contrast, first movers who took money from second movers made an average profit of only 0.15 euro after the second movers made their decisions. Next consider the behavior of second movers. Note that 13 of the 30 second movers neither gave nor took money from first movers. But 17 second movers did reduce their own money payoffs in order to either give or take

money from first movers; five of them took money from first movers and 12 gave money to second movers.

4.1. Behavior in Our Moonlighting Game vs. the Abbink, et al. Moonlighting Game

First- and second-mover data from our (moonlighting game) treatment A can be compared with data from the "without contracts" moonlighting game treatment reported in Abbink, Irlenbusch, and Renner (2000). In their experiment, each subject was given a credit balance of 12 "talers," the fictitious currency of the experiment. First movers could send and second movers return integer amounts. Positive amounts sent by first movers were tripled by the experimenters, while negative amounts "sent" were not transformed. A first mover could send an amount that varied from –6 to +6 (or from –6 to +18, including the tripling of positive amounts). Negative amounts "returned" were tripled by the experimenters, while positive amounts returned were not transformed. A second mover could return an amount that varied from –6 to +18 (or from –18 to +18, including the tripling of negative amounts).

In the Abbink, et al. experiment, two out of the 32 first movers took the maximum of six talers and four others took smaller amounts from the paired second mover. Five subjects "sent" zero and 21 subjects gave positive amounts to the second mover; ten subjects gave the maximum of 18 talers. On average, it was not profitable (in monetary terms) for the first movers to either give or take money. But the average loss from taking was larger than the average loss (in monetary terms) from giving money. First movers who gave positive amounts of money to second movers made an average loss of 0.3 talers. First movers who took money from second movers made an average loss of 4.2 talers (or, alternatively, an average loss of 7.4 talers if one excludes the observation excluded by Abbink, et al.).[7]

Next consider the behavior of second movers. In their experiment, 12 of the 32 second movers neither gave nor took money from first movers. But 20 second movers did reduce their

own money payoffs in order to either give or take money from first movers; six of them took money from first movers and 14 gave money to second movers.

Table 1 presents contingency tables comparing first- and second-mover data from the Abbink, et al. without-contracts treatment with data from our treatment A. Chi-square contingency table tests are used because the two experiments involve different strategy sets. The contingency table for amounts sent by first-movers includes two classes: class 1 is the number of observations for which the amount sent, s was nonpositive and class 2 is the number of observations for which the amount sent was positive. The test reveals no significant difference ($p > .1$) between first-mover data from the two experiments. The contingency table for amounts returned, r by second-movers includes the three classes for which the amounts returned are zero, positive, or negative. It is natural to separate observations of $r = 0$ from both positive and negative values of r because a return of zero is the unique prediction of the self-regarding preferences model. The test reveals no significant difference ($p > .8$) between second-mover data from the two experiments.

4.2 Tests of the Self-Regarding Preferences Model

We first consider subjects' behavior in treatment A. The traditional model of self-regarding preferences has straightforward predictions for this game. Since a second mover would be assumed to want only to maximize her own money payoff, she would be predicted neither to take nor give money to the paired first mover because either action would be costly to her. Knowing this, and assumed only to want to maximize his own money payoff, a first mover would be predicted to take the maximum of five euro (i.e. send minus five) from the paired second mover.

As noted above, 12 out of the 30 first movers took five euro from their paired second movers and 13 out of the 30 second movers neither gave nor took money from first movers. Aggregating over all subjects in treatment A, 35 out of 60 or 58% of the subjects made decisions that are inconsistent with the self-regarding preferences model. Focusing on pairs of subjects,

we observe that six of the first movers who took the maximum of five euro were not punished by their paired second movers. Thus the behavior of six out of 30 or 20% of the subject pairs is consistent with the subgame perfect equilibrium of the self-regarding preferences model.

The first row of Table 2 reports the means and standard deviations of the amounts sent by first movers and returned by second movers in treatment A. The fourth row of the table reports results from a Kolmogorov test of the hypothesis that amounts sent are equal to minus five; the test implies rejection of the hypothesis. The Kolmogorov test reported in the fifth row of Table 2 implies rejection of the hypothesis that amounts returned are equal to zero. We conclude that subjects' behavior in treatment A is not consistent with the predictions of the self-regarding preferences model.

We next describe the subjects' behavior in the dictator treatments. The subjects' behavior in treatment B is presented in Figure 6 and their behavior in treatment C is shown in Figure 7. The self-regarding preferences model predicts that the maximum of five will be taken by "first movers" in treatment B and that zero will be returned by "second movers" in treatment C. In treatment B, 21 out of 27 subjects took the maximum of five euro, four subjects took smaller amounts, and two subjects gave positive amounts. In treatment C, 21 out of 30 subjects chose zero euro, six subjects "returned" positive amounts, and three subjects "returned" negative amounts. The second and third rows of Table 2 report the means and standard deviations of the amounts sent and returned in treatments B and C. The Kolmogorov test reported in row six does not imply rejection of the hypothesis that amounts sent in treatment B are equal to minus five. The Kolmogorov test reported in the seventh row of Table 2 does not imply rejection of the hypothesis that amounts returned are equal to zero. Thus the tests do not reject the predictions of the self-regarding preferences model with data from treatments B and C. In treatment A, 42% of the subjects made decisions that are consistent with the self-regarding preferences model. In contrast, in treatments B and C, 42 out of 57 or 74% of the subjects made decisions that are consistent with the model's predictions.

4.3 Tests for Trust, Fear, Altruism, and Reciprocity

The implications of the self-regarding preferences model are inconsistent with the data for the majority of subjects in treatment A. Thus the subjects have motivations that are richer and more complicated than simply a desire to maximize their own money payoffs in the experiment. We next examine the information about alternative motivations that is provided by the triadic experimental design.

First consider the behavior of first movers. Figure 6 presents the amounts sent in treatments A and B. The third row of Table 3 reports tests comparing first-mover behavior in treatments A and B. All three of the reported tests imply the conclusion that there is a highly-significant difference between first-mover sending behavior in treatments A and B. Thus first movers behave quite differently when the second movers have an opportunity to respond than when they do not.

Altruism could motivate sending positive amounts in either treatment A or B. In contrast, first movers' trust that second movers will not defect could motivate the sending of positive amounts in treatment A but not in treatment B. The experimenters triple any positive amounts sent by first movers. This creates a monetary profit that can be shared between first and second movers in treatment A if the second movers do not defect. Furthermore, first movers' fear of the negative reciprocity and/or inequality-averse preferences of second movers could motivate them to avoid taking money in treatment A but not in treatment B. Thus, comparison of subjects' behavior in treatments A and B permits one to discriminate among some alternative motivations.

Altruism is the only motive for first movers to send positive amounts in treatment B. As noted above, 25 out of 27 subjects took money in treatment B. Only two of the subjects exhibited altruistic motivation by sending positive amounts in treatment B. The last row of Table 2 reports tests of the hypothesis that amounts sent in treatment B are greater than or equal to zero. Not surprisingly, the tests imply rejection of the hypothesis that amounts sent are non-negative. We conclude that subjects' behavior in treatment B is not characterized by significant altruism. Since

the treatment A subjects are drawn from the same subject pool as the treatment B subjects, we conclude that their behavior is also not characterized by significant altruism.

A selfish first mover might send a non-positive amount in treatment B but send a positive amount in treatment A because of trust that the second mover would share the monetary profit from the tripling of amounts sent. As seen in Figure 6, 14 out of 30 first movers sent positive amounts of money to second movers in treatment A. In contrast, only two first movers sent positive amounts of money to second movers in treatment B. We conclude that many first movers exhibited trust in the moonlighting game. If we base our figure on the non-rejection of the hypothesis that subjects sent minus 5 in treatment B, we conclude that 47% (= (14/30) × 100) of the subjects in treatment A exhibited trust. Alternatively, if we base our figure on the difference between the fractions of subjects that sent positive amounts in treatment A (14/30) and treatment B (2/27), we conclude that 39% (=(14/30-2/27) × 100) of the subjects in treatment A exhibited trust.

A test for the significance of trusting behavior can be constructed as follows. If $s_a > 0$ and $s_a > s_b$ then there is evidence of trusting behavior in the moonlighting game (appendix Proposition 6.1.(b)). Let us equivalently rewrite it as $s_a > \max\{s_b, 0\}$ implies a trusting action. Create a new random variable, x as follows:

(10) $x = 0$, if $s_b < 0$

$= s_b$, otherwise.

Let F_S be the cumulative distribution function for s_a and F_X be the cumulative distribution function for x. The null hypothesis is that s_a and x have the same distribution and the alternative hypothesis is that s_a is strictly stochastically larger than x (by first-order stochastic dominance):

$$H_0^T: \ F_S(y) = F_X(y)$$

$$H_a^T: \ F_S(y) < F_X(y)$$

The one-tailed Kolmogorov-Smirnov two-sample test has a *p*-value in the interval (0.01,0.02). Therefore, we conclude that the first movers' behavior in the moonlighting game is characterized by significant trust since $F_S(y) < F_X(y)$ implies the sufficient condition for trust.

Next consider the question of whether the significantly higher amounts sent by first movers in treatment A than in treatment B are motivated by fear or trust. A first mover might prefer to take money from the paired second mover. If so, he will take money in treatment B, but may also refrain from taking money in treatment A if he is afraid of retaliation by the second mover due to negative reciprocity and/or inequality-averse preferences, or he may send money in treatment A if he trusts in positive reciprocity and/or altruistic preferences. A selfish, fearful first mover would send zero in treatment A and take five in treatment B. A selfish, unafraid first mover would take money in both treatments A and B. A selfish, trusting first mover would send positive amounts of money in treatment A and take five in treatment B. As seen in Figure 6, three out of 30 of the subjects chose zero in treatment A, whereas 25 out of 27 of the subjects took money in treatment B. Thus the behavior of only 11% (= ((3/30) ÷ (25/27)) × 100) of the subjects is consistent with fear of negative reciprocity. One also observes from Figure 6 that 13 out of 30 of the first movers in treatment A took money from the paired second mover. Thus the behavior of 47% (= ((13/30) ÷ (25/27)) × 100) of the subjects is inconsistent with fear of negative reciprocity.

A test for the significance of fearful behavior can be constructed as follows. Let A_{NP} be the subset of observations from treatment A with s_a nonpositive. Those data can be generated from subjects that do not exhibit trust because $s_a \leq 0$ is a sufficient condition for the absence of trust (see appendix Proposition 6.1.(a)). Let n denote the size of A_{NP}. Order the nonpositive data from treatment B from the lowest to the highest and let B_n be the first n observations in the

above ordered sequence. Note that $s_{b,i} \leq 0, i \leq n$. Let F_A be the empirical cumulative distribution function of the observations in A_{NP} and let F_B be the empirical cumulative distribution function of the observations in B_n. The null hypothesis is that $s_{A_{NP}}$ and s_{B_n} are drawn from the same distribution and the alternative hypothesis is that $s_{A_{NP}}$ is strictly stochastically larger than s_{B_n}:

$$H_0^T: \ F_A(y) = F_B(y)$$

$$H_a^T: \ F_A(y) < F_B(y)$$

The one-tailed Kolmogorov-Smirnov two-sample test has a *p*-value in the interval (0.5,0.3). Hence, the null hypothesis of $F_A(y) = F_B(y)$ for the selected data is not rejected. Since $F_A(y) < F_B(y)$ implies the necessary and sufficient condition for fear (derived in appendix Proposition 6.2), we conclude that the selected data are not characterized by fear. Furthermore, note that B_n contains observations from treatment B with the lowest values of s_b, and therefore it favors the alternative hypothesis of fear. Thus if the hypothesis of "no fear" is not rejected in A_{NP} versus B_n it cannot be rejected in A_{NP} versus any nonpositive subset of size n of B. Therefore we conclude that the first movers' behavior, that is not characterized by trust, is characterized by an absence of fear.

We now consider the behavior of second movers. A "second mover" in treatment C has the same strategy set as a second mover in treatment A. The allocated money payoffs of the first and second movers, prior to the second mover's decision, are the same in treatments A and C. The difference between the treatments is that first movers' decisions determine these allocations in treatment A but not in treatment C. Thus, second movers can be motivated by reciprocity in treatment A but not in treatment C. Whether or not the behavior of second movers *is* characterized by reciprocity is revealed by comparing responses in treatments A and C. Figures 5

and 7 show how second movers responded to amounts they received in treatments A and C. Figure 8 presents a direct comparison of amounts returned in treatments A and C. We first consider responses by second movers who received positive amounts.

Fourteen second movers received positive amounts of money sent by the paired first movers in treatment A and were provided correspondingly-larger endowments by the experimenters in treatment C. How did they respond in each of the two treatments? In treatment A, 11 responded by returning positive amounts to first movers and three second movers kept all of the money. In contrast, in treatment C three second movers returned positive amounts to first movers and 11 second movers kept all of the money. Another striking difference between the treatments is for the five second movers in each treatment who received the maximum of 30 euro. In treatment C, all five of such second movers kept all of the money. In contrast, in treatment A all of them returned positive amounts, with the amounts returned varying from a low of 10 euro to a high of 20 euro. Finally, note that the fourth row of Table 3 reports tests comparing amounts returned in treatments A and C by second movers who received positive amounts. All of the tests detect a highly significant difference between the treatments. We conclude that the behavior of subjects in the moonlighting game is characterized by significant positive reciprocity.

Next consider responses by the 13 second movers who "received" negative amounts in both treatments. 12 of these subjects had the maximum amount of five euro taken from them and the other subject had one euro taken in treatment A. Corresponding endowments were given in treatment C. How did the subjects respond in each of the two treatments? In treatment A, five second movers responded by incurring a cost to take money from the paired first mover, seven responded by choosing zero, and one responded by giving the first mover one euro. The behavior of the five second movers who took money in treatment A could be explained by either negative reciprocity or inequality aversion. In treatment C, three second movers responded by incurring a cost to take money from the paired "first mover," eight responded by choosing zero, and two responded by giving the "first mover" one euro. The behavior of the three second movers who

took money in treatment C could be explained by inequality aversion but *not* by negative reciprocity. Thus, whether or not the behavior of second movers *is* characterized by negative reciprocity is revealed by comparing responses in treatments A and C. The last row of Table 3 reports tests comparing amounts returned in treatments A and C by second movers who received negative amounts. The tests do not detect a significant difference. We conclude that the behavior of subjects in the moonlighting game is not characterized by significant negative reciprocity since $r_a < \min\{r_c, 0\} = r_c$, for $s_a < 0$, is a necessary and sufficient condition for negative reciprocity. (See appendix Proposition 6.4.)

5. Further Exploration of Subjects' Behavior

Differences between subjects' behavior in the moonlighting game and their behavior in the control dictator games are what imply the conclusions that subjects' behavior is characterized by trust and positive reciprocity and not characterized by fear or negative reciprocity. Inspection of Figures 4-7 makes clear that subjects' behavior in the dictator games was generally not characterized by altruism, as distinct from behavior in many other dictator games. What accounts for the unusually selfish behavior in our dictator games?

In a typical dictator game, the "second mover" will not be given an endowment. The first mover will be given an endowment, perhaps $10. In that case, the average amount given by the dictators is typically $1 or $2, depending on specific features of the experiment protocol. For example, in the (DB1 and DB2) double blind dictator experiments reported by Hoffman, McCabe, Shachat, and Smith (1994), the average amount sent to the paired subjects by the dictators was $1, leaving first movers with average payoff of $9 and "second movers" with average payoff of $1. A more striking example of altruism was observed in the first mover dictator control treatment of the triadic experimental design for the investment game (Cox, 2002b). In that dictator game treatment, first movers gave second movers, on average, $3.63.

With the multiplication by three, and recalling that both first and second movers got \$10 endowments, first movers' average payoff was \$6.37 and second movers' average payoff was \$20.89 in treatment B for the investment game triad. In sharp contrast, the average amount transferred by first movers in treatment B in the present, moonlighting game triad is negative, in fact $-\$3.11$. This produces average first mover payoff of \$13.11 and average second mover payoff of \$6.89.

The unexpected selfishness of subjects in the moonlighting-triad treatment B caused us to question the data. We wondered if there was some unusual feature of the computerized protocol or the University of Amsterdam subject pool that produced anomalous behavior in the moonlighting-triad dictator control treatments. In order to check on this, one experimenter (Cox) used the same *manual* (i.e., non-computerized) protocol and University of Arizona subject pool used in the earlier investment-triad experiment (Cox, 2002b) to replicate the moonlighting-triad treatment B. This experiment produced data that are not significantly different from the treatment B data used in this paper.

Further reflection suggests that one should not be surprised by the dictator game data, after all. Suppose that subjects have preferences characterized by utility functions that are globally increasing in both one's own money payoff and the other person's money payoff. Further suppose that the other-regarding preferences are "egocentric," which means that if $x > y$ then I prefer the outcome in which I get \$$x$ and you get \$$y$ to the outcome in which I get \$$y$ and you get \$$x$ (Cox, Sadiraj, and Sadiraj, 2002). Indifference curves for such a preference ordering, that rationalize the seemingly anomalous data, are shown in Figure 9. Point T is the average observation in a typical, split-the-\$10 dictator game. Point I is the average observation in the investment-triad treatment B dictator control game in which the dictator's budget line consists only of the line segment that extends from the 45-dgree line to the vertical axis. And point M is the average observation in the moonlighting-triad treatment B dictator control game in which the dictator's piece-wise-linear "budget line" includes the line segments above and below the 45-

degree line. Thus, upon reflection the data are not anomalous; instead, they suggest a model that can be used to rationalize data from many different types of experiments for which subjects' intentions are not significant determinants of behavior (Cox, Sadiraj, and Sadiraj, 2002).

6. Concluding Remarks

This paper reports an experiment with a game triad that includes the moonlighting game. Abbink, Irlenbusch, and Renner (2000) had previously reported data for the moonlighting game that are consistent with reciprocity but inconsistent with the traditional self-regarding preferences model. These results, and results from many other non-market fairness experiments (Fehr and Gächter, 2000), leave the profession with the task of constructing alternatives to the self-regarding preferences model in order to gain consistency with the empirical evidence. But this task cannot be undertaken successfully unless we can discriminate among the observable implications of alternative possible motivations. The game triad experiment reported here makes it possible to discriminate among the observable implications for subjects' choices of trust, fear, reciprocity, and altruism in the moonlighting game.

Results from our experiment support the conclusion that 39% to 47% of first movers in the moonlighting game were motivated by trust that the second mover would not defect. Furthermore, this trust was based on realistic expectations because the behavior of second movers who received positive amounts from first movers was characterized by significant positive reciprocity. Indeed, positive reciprocity caused trusting behavior to have positive expected profit: first movers who sent positive amounts to second movers made an average profit of 1.93 euro after the second movers' decisions. The behavior of 47% of the first movers is inconsistent with fear of negative reciprocity. This absence of fear was based on realistic expectations because the behavior of second movers who had money taken from them by first movers was not characterized by significant negative reciprocity. Indeed, the absence of significant negative

reciprocity caused taking behavior to have a small positive expected profit: first movers who took money from second movers made an average profit of 0.15 euro after the second movers' decisions. But this conclusion about the realistic expectations of first movers who took money needs some qualification because the average profit of first movers who gave money to second movers was notably higher than the average profit of those who took money from them.

Falk, Fehr, and Fischbacher (2001) report experiments with a design that includes the moonlighting game and a control treatment in which the first-mover amounts taken or sent to the second movers are randomly generated. Their design uses the "strategy method" in which second movers choose responses to all possible first-mover decisions, or random determinations of first-mover amounts, before observing the actual amounts taken or sent by first movers. Another way in which their design differs from ours is that they use a single-blind payoff protocol in which individual subjects' decisions are known by the experimenters.[8] They conclude that their subjects' behavior is characterized by both positive and negative reciprocity.

It is presently unclear which of the differences between the Falk, et al. experimental design and our experimental design accounts for the different conclusions about the significance of negative reciprocity. But both experiments generate data that support the conclusion that attributions of intentions are a significant determinant of behavior in the moonlighting game. Thus both experiments lead to the conclusion that behavior in the moonlighting game cannot be fully explained by models of preferences over outcomes, such as models of inequality aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) and egocentric other-regarding preferences (Cox, Sadiraj, and Sadiraj, 2002). The two experiments support the conclusion that models that include both intentions and preferences over outcomes are needed to fully explain behavior in the moonlighting game. Models that include both intentions and outcome preferences have been developed by Falk and Fischbacher (1999), Charness and Rabin (forthcoming), and Cox and Friedman (2002).

Papers reporting experiments with some other games also lead to the conclusion that both intentions and outcome preferences are needed in models. These other games include the ultimatum game (Blount, 1995), the investment game (Cox, 2002a, 2002b), and the mini-ultimatum game (Falk, Fehr, and Fischbacher, 1999). Some other experiments yield mixed results. Bolton, Brandts, and Ockenfels (1998) report that intentions are an insignificant determinant of behavior. Charness (2001) and Offerman (1999) report that intentions are significant for negative reciprocity but not for positive reciprocity. Cox and Deck (2002) find that negative reciprocity is not significant in the punishment mini-ultimatum game and that positive reciprocity is significant in the trust mini-investment game with a single-blind protocol but insignificant with a double-blind protocol.

We conclude that behavior in games involving salient fairness considerations is rich, complicated, and difficult fully to explain with game-theoretic modeling. But it is quite clear that the "economic man" model of simple self-regarding preferences is not sufficiently rich to explain economic behavior in fairness games. Successful models must not only incorporate other-regarding preferences, but also preferences that are conditional on the perceived intentions of others. The experimental evidence supporting this characterization of behavior is of recent vintage. But the characterization itself is very old:

> *Before any thing, therefore, can be the complete and proper object, either of gratitude or resentment, it must possess three different qualifications. First it must be the cause of pleasure in the one case, and of pain in the other. Secondly, it must be capable of feeling these sensations. And, thirdly, it must not only have produced these sensations, but it must have produced them from design, and from a design that is approved of in the one case and disapproved of in the other. – Adam Smith (1759, p. 181).*

Endnotes

* Financial support was provided by the Center for Research in Experimental Economics and Political Decisionmaking (CREED), University of Amsterdam and by the Decision Risk and Management Science Program, National Science Foundation (grant number SES-9818561). We are grateful to Ingrid Seinen for help in writing the subject instructions and text material for the screen displays in Dutch. Jos Theleen, the CREED programmer did a fine job in developing the software.

1. Models of inequality-averse preferences are developed and applied in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000).
2. Models of (unconditional) other-regarding preferences over outcomes are presented in Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Andreoni and Miller (forthcoming), and Cox, Sadiraj, and Sadiraj (2002). Models of intentions are presented in Rabin (1993) and Dufwenberg and Kirchsteiger (1999). Models of intentions and outcome preferences are presented in Falk and Fischbacher (1999) and Charness and Rabin (forthcoming), and Cox and Friedman (2002).
3. In writing this paper, we follow the convention officially requested by the European Union: that, in the case of the euro, an exception be made to the usual English language distinction between singular and plural nouns. The EU has requested that "euro" be used for one and many currency units.
4. The $10 endowment and $1 unit of divisibility experimental design was used by Berg, et al. (1995), Croson and Buchan (1999), and Cox (2002a, 2002b).
5. The subject instructions are available online at http://uaeller.eller.arizona.edu/~jcox/index.html.

6. See Dufwenberg, Gneezy, Güth, and van Damme (2001) for tests of both direct and indirect reciprocity in the context of the investment game.
7. Abbink, Irlenbusch, Renner (2000, fn. 10) state that they excluded from their data analysis the observation shown in their Figure 2 for which the second mover gave his entire remaining balance of eight talers to the paired first mover who took four talers from him. They do not provide an explanation of their decision to exclude the observation.
8. The evidence is mixed concerning whether the level of social distance in a protocol is a significant determinant of behavior in fairness experiments. Hoffman, et al. (1994) and Cox and Deck (2002) found significant effects from the single-blind/double-blind treatment while Bolton and Zwick (1995), Bolton, Katok, and Zwick (1998), and Johanneson and Persson (2000) did not.

References

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

Andreoni, James and J. Miller, "Giving According to GARP: An Experimental Test of the Rationality of Altruism," forthcoming in *Econometrica*.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, July 1995, 10(1), pp. 122-42.

Blount, Sally, "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes*, August 1995, 63(2), pp. 131-44.

Bolton, Gary, Jordi Brandts, and Axel Ockenfels, "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game." *Experimental Economics*, 1998, 1(3), pp. 207-19.

Bolton, Gary E., Elena Katok, and Rami Zwick, "Dictator Game Giving: Rules of Fairness versus Acts of Kindness." *International Journal of Game Theory*, 1998, 27, pp. 269-99.

Bolton, Gary E. and Axel Ockenfels, "ERC: A Theory of Equity, Reciprocity and Competition." *American Economic Review*, March 2000, 90(1), pp. 166-93.

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

Charness, Gary, "Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation." Working paper, University of California at Berkeley, September 1996, revised April 2001.

Charness, Gary and Matthew Rabin, "Social Preferences: Some Simple Tests and a New Model." Forthcoming in the *Quarterly Journal of Economics*.

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

Cox, James C., "How to Identify Trust and Reciprocity." Discussion paper, University of Arizona, 2002b.

Cox, James C. and Cary A. Deck, "On the Nature of Reciprocal Motives." Discussion paper, University of Arizona, September 2000; revised 2002.

Cox, James C. and Daniel Friedman, "A Tractable Model of Reciprocity and Fairness," Discussion Paper, University of Arizona, 2002.

Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj, "A Theory of Competition and Fairness for Egocentric Altruists," Discussion paper, University of Arizona and University of Amsterdam, January 2001; revised 2002.

Croson, Rachel and Nancy Buchan, "Gender and Culture: International Experimental Evidence from Trust Games." *American Economic Review*, 1999, 89, pp. 386-91.

Dufwenberg, Martin, Uri Gneezy, Werner Güth, Eric van Damme, "Direct versus Indirect Reciprocity: An Experiment." *Homo Oeconomicus*, 2001, 18, pp. 19-30.

Dufwenberg, Martin and Georg Kirchsteiger, "A Theory of Sequential Reciprocity." Discussion paper, CentER for Economic Research, Tilburg University, 1999.

Falk, Armin, Ernst Fehr, and Urs Fischbacher, "Testing Theories of Fairness – Intentions Matter." University of Zurich discussion paper, May 2001.

Falk, Armin and Urs Fischbacher, "A Theory of Reciprocity." Working Paper No. 6, Institute for Empirical Research in Economics, University of Zurich, 1999.

Fehr, Ernst and Simon Gächter, "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, Summer 2000b, 14(3), pp. 159-81.

Fehr, Ernst and Klaus M. Schmidt, "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, August 1999, 114(3), pp. 817-68.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon L. Smith, "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 1994, 7, pp. 346-80.

Johanneson, M. and B. Persson, "Non-reciprocal Altruism in Dictator Games." *Economics Letters*, 2000, 69, pp. 137-42.

Offerman, Theo, "Hurting Hurts More than Helping Helps: The Role of the Self-Serving Bias." Working paper, University of Amsterdam, 1999.

Rabin, Matthew, "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 1993, 83, pp. 1281-1302.

Smith, Adam, *The Theory of Moral Sentiments*, 1759; reprinted by Indianapolis: Liberty Classics, 1976.

Table 1. Treatment A Data vs. Abbink, et al. Data

	Send Data Contingency Table		Return Data Contingency Table			
	Class 1 ($s \leq 0$)	Class 2 ($s > 0$)	Class 1 ($r = 0$)	Class 2 ($r < 0$)	Class 3 ($r > 0$)	
Treatment A	$16_{13.07}$	$14_{16.94}$	$12_{12.90}$	$6_{5.68}$	$14_{13.42}$	
Abbink, et al.	$11_{13.94}$	$21_{18.07}$	$13_{12.10}$	$5_{5.32}$	$12_{12.58}$	
Chi-Square Test	2.26 (p>.1)		0.22 (p>.8)			

Table 2. Tests of Predicted Distributions

Data	Send Mean	Return Mean	Kolmogorov Test
Tr. A	7.47 [13.88] {30}	2.10 [9.02] {30}	…
Tr. B	-3.11 [6.83] {27}	…	…
Tr. C	…	-0.20 [2.91] {30}	…
Tr. A Send vs. Minus Five	…	…	0.60 (p <.01)[1]
Tr. A Ret. vs. Zero	…	…	0.40 (p <.01)
Tr. B Send vs. Minus Five	…	…	0.22 (.05<p<.1)[1]
Tr. C Ret. vs. Zero	…	…	0.20 (p=.15)
Tr. A Send vs. Zero	…	…	0.43 (p <.005)[1]
Tr. B. Send vs. Zero	…	…	0.93 (p <.005)[1]

Standard deviations in brackets.
Number of subjects in braces.
p-values in parentheses.
[1] denotes a one-tailed test.

Table 3. Tests for Trust, Fear, and Reciprocity

Data	Return Mean			Means Test (eq. var.)	Smirnov Test	Mann-Whitney Test
	Send A < 0	Send A > 0	All Send A			
Tr. A	-4.54 [6.84] {13 }	8.71 [6.78] {14}	2.10 [9.02] {30}	…	…	…
Tr. C	-1.46 [3.48] {13}	0.93 [2.16] {14}	-0.20 [2.90] {30}	…	…	
Tr. A Send vs. Tr. B Send (trust and/or fear)	…			3.59 (.000)[1]	0.53 (p<.005)[1]	3.33 (p<.001)[1]
Tr. A Return vs. Tr. C Return (send A > 0)	…			4.10 (.000)[1]	0.71 (p<.005)[1]	3.25 (p<.001)[1]
Tr. A Return vs. Tr. C Return (send A < 0)	…			-1.45 (.081)[1]	0.31 (p>0.1)[1]	-1.20 (.115)[1]

Standard deviations in brackets.
Number of subjects in braces.
p-values in parentheses.
[1] denotes a one-tailed test.

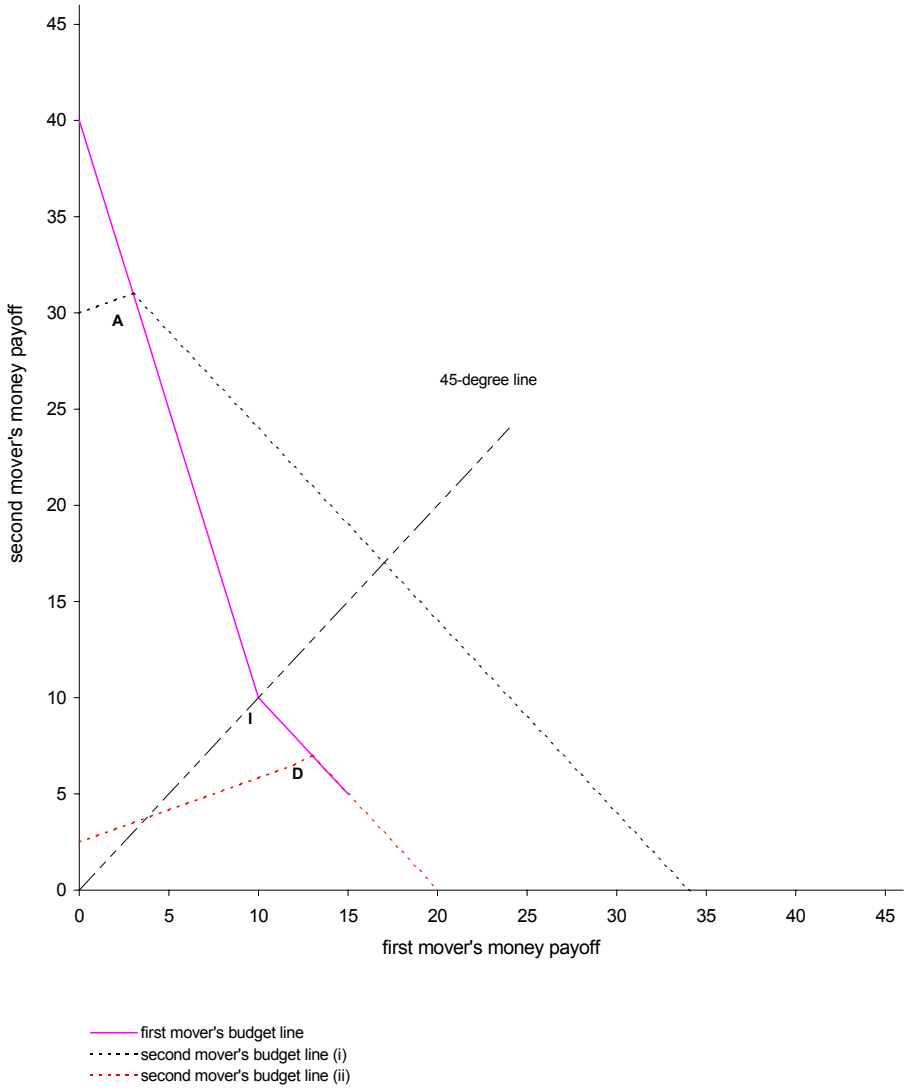

Figure 1. Illustration of Representative Budget Lines of First and Second Movers

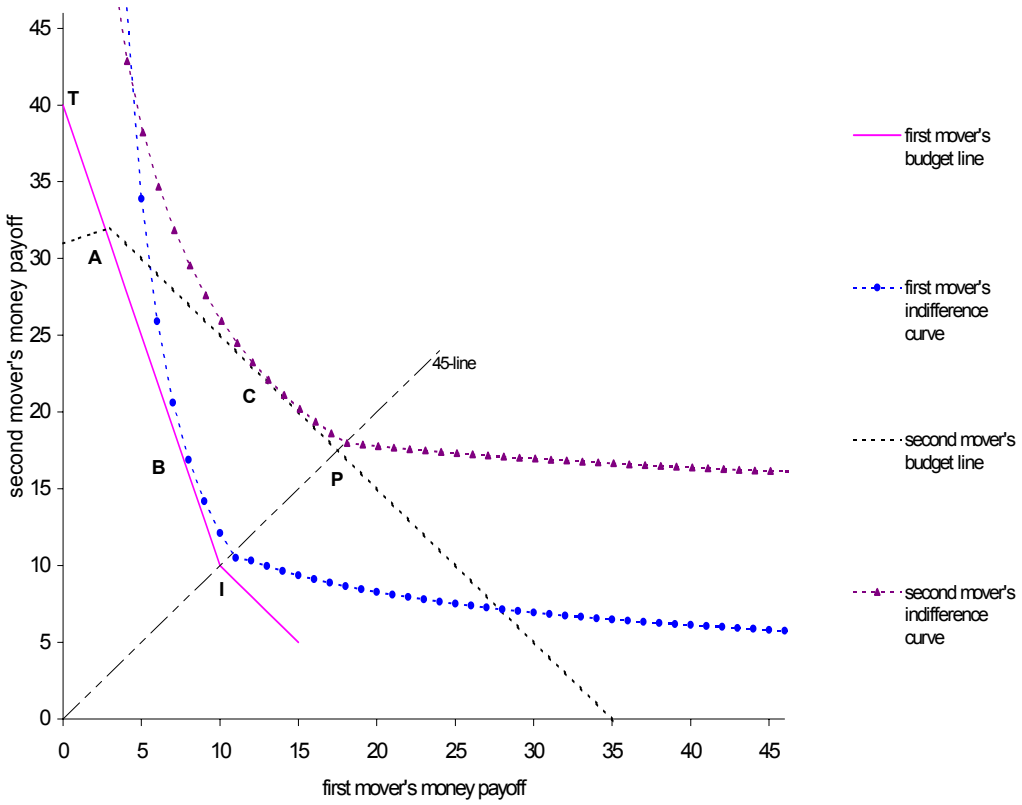

Figure 2. Illustration of Trust and Positive Reciprocity

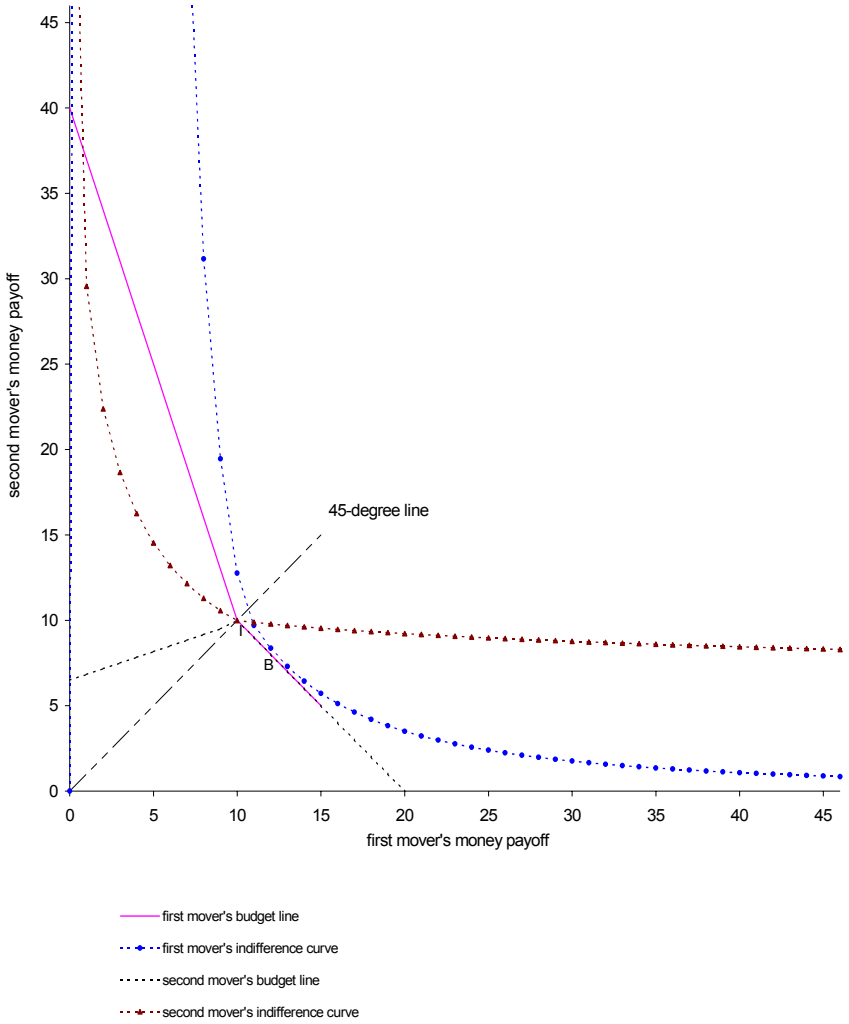

Figure 3. Illustration of Fear and No Reciprocity

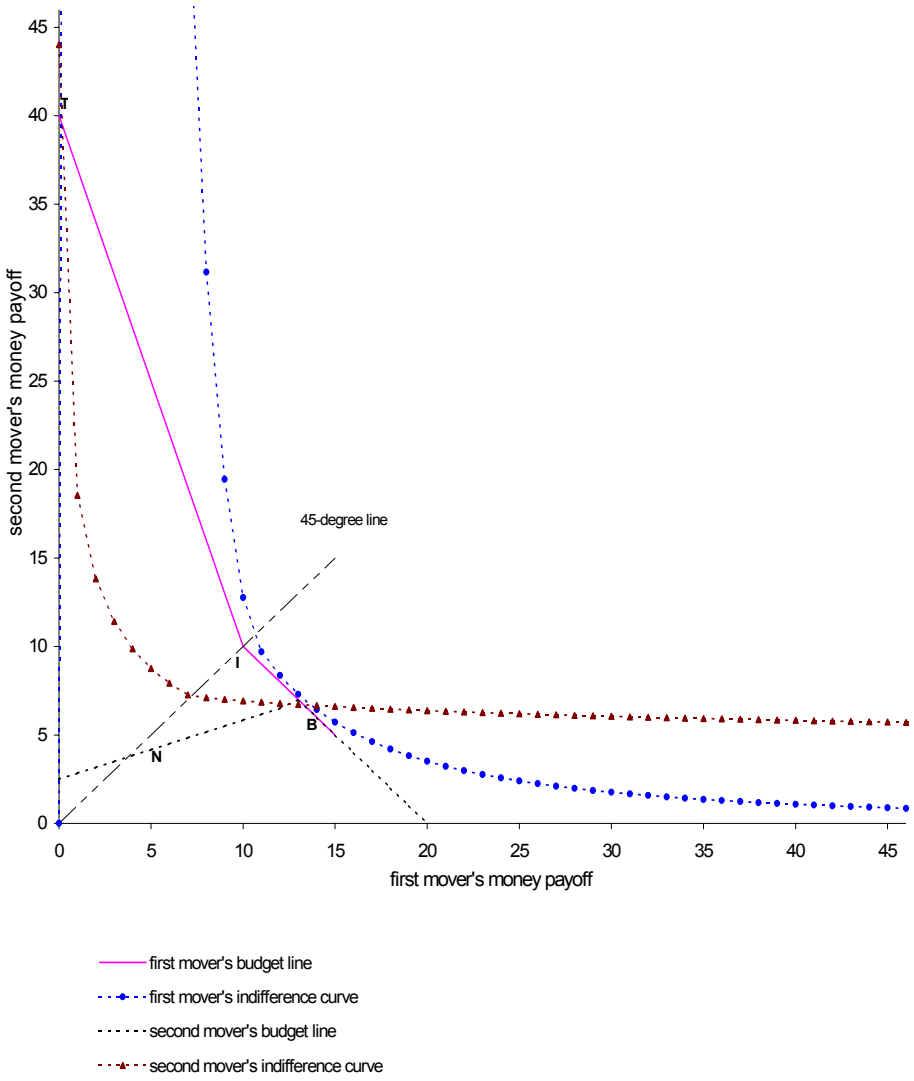

Figure 4. Illustration of No Fear and Negative Reciprocity

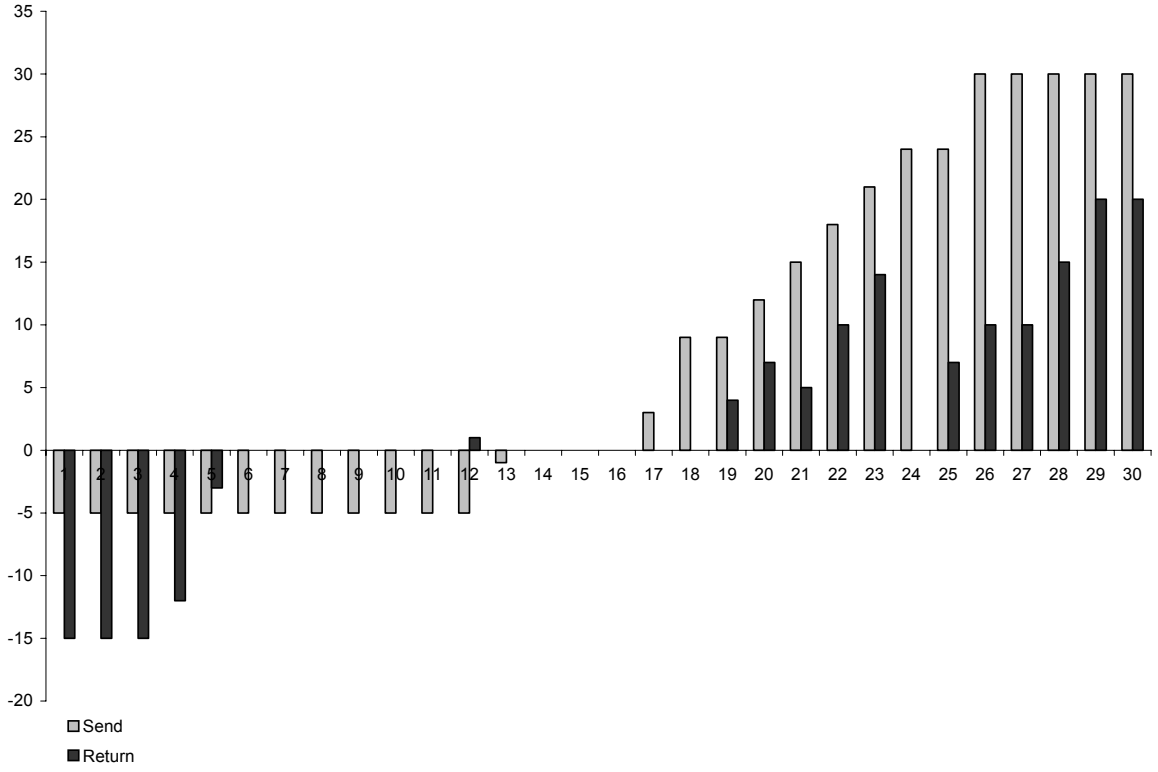

Figure 5. Money Sent and Returned in Treatment A

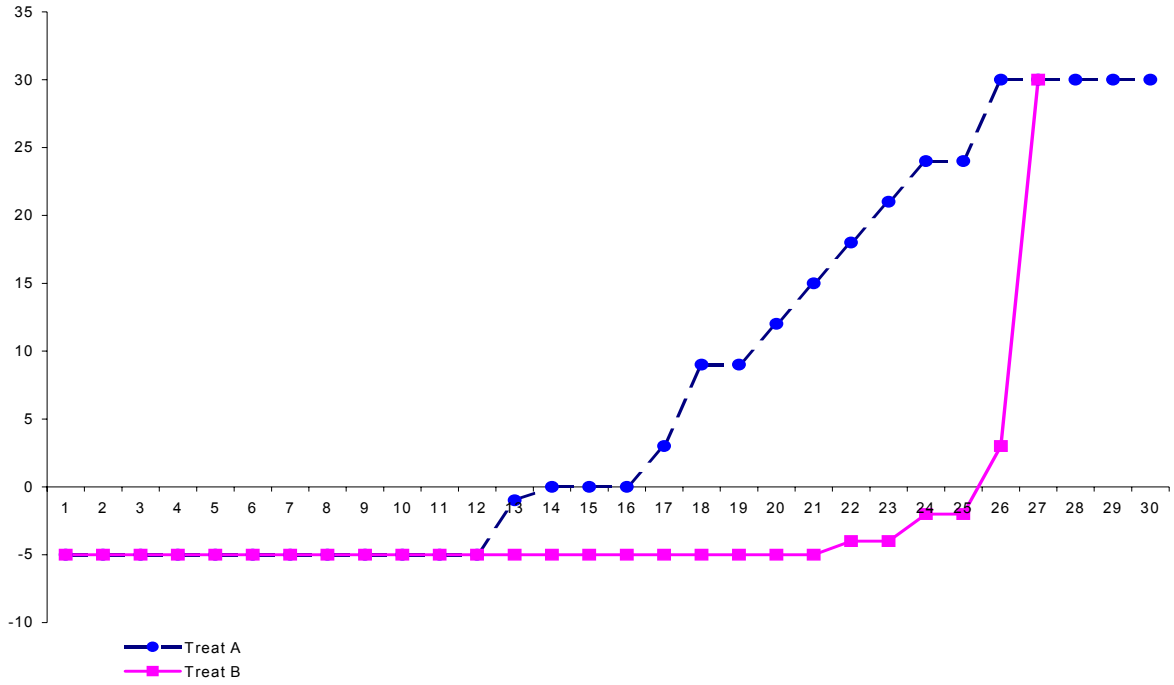

Figure 6. Money Sent in Treatments A and B

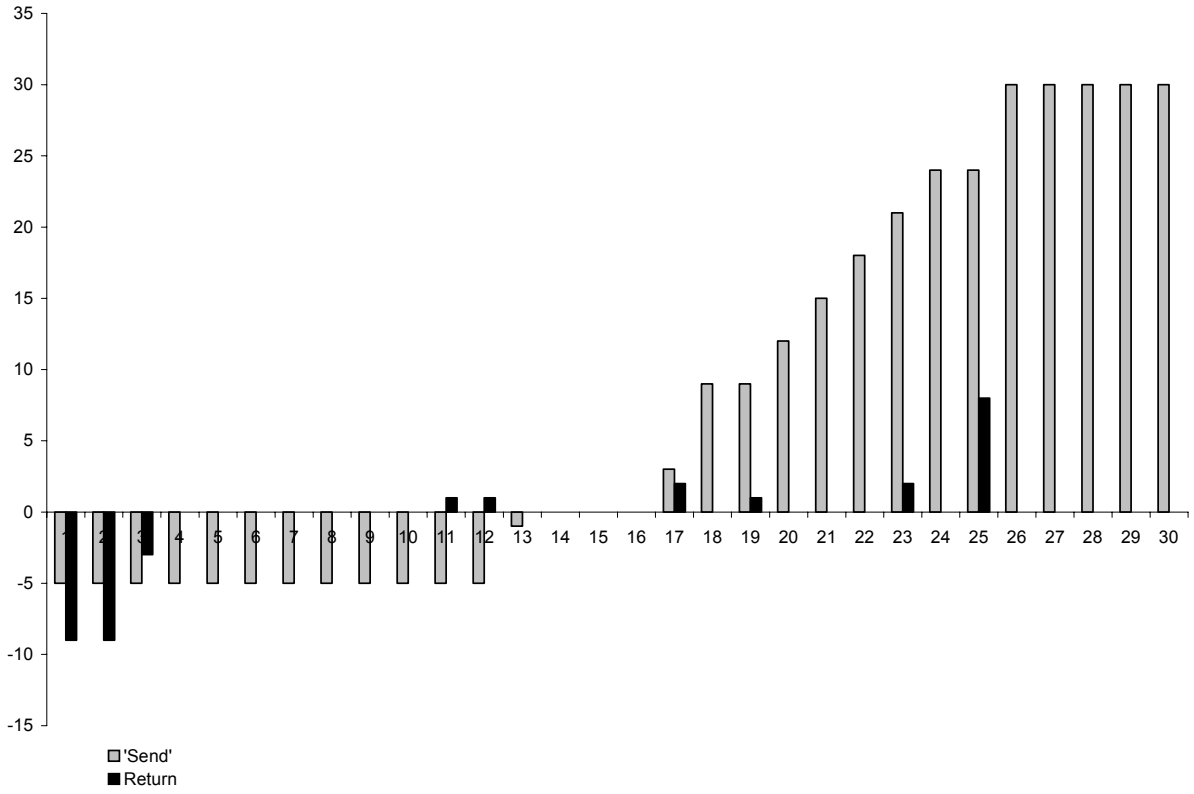

Figure 7. Money "Sent" and Returned in Treatment C

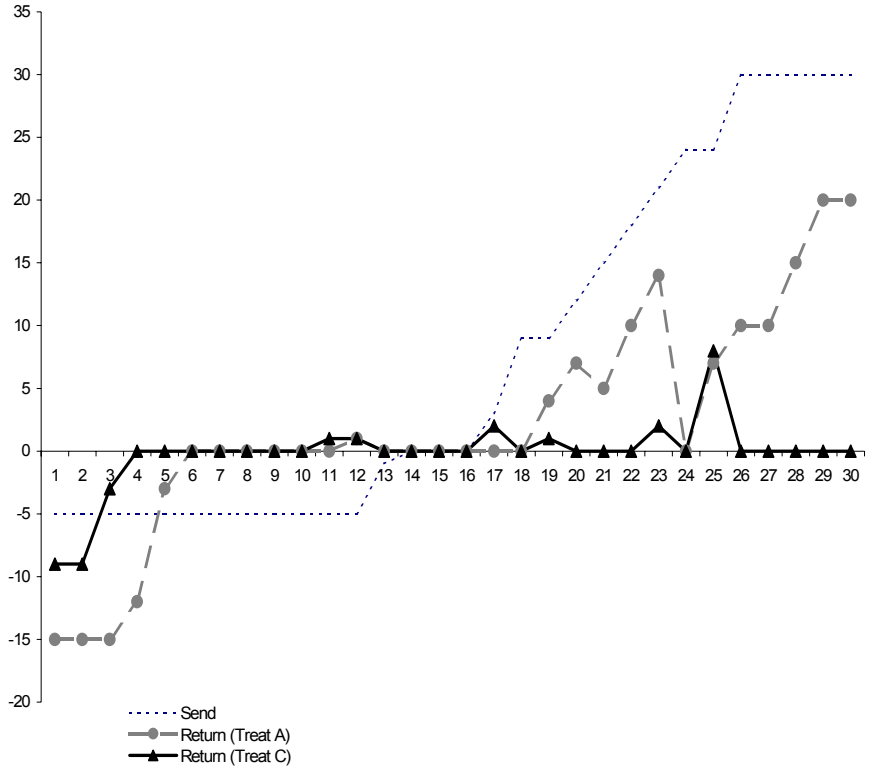

Figure 8. Money Sent in Treatment A and Returned in Treatments A and C

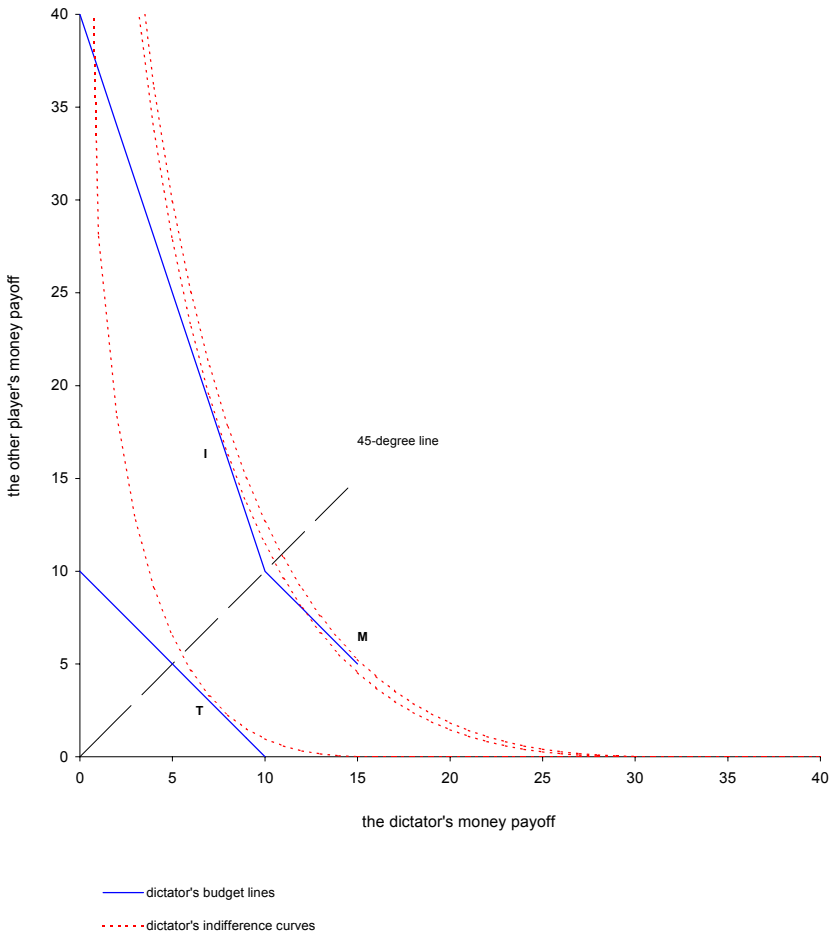

Figure 9. Rationalizing Data from Various Dictator Games

Appendix: Derivations of Test Criteria for Trust, Fear, and Reciprocity

Definitions

The concepts of trust and positive reciprocity used in this paper are defined as follows.

Definition 1 *A first mover undertakes an action that exhibits* ***trust*** *if the chosen action:*

(a) *sends the second mover a positive amount of money, which creates a monetary gain that could be shared; and*

(b) *exposes the first mover to the risk of a loss of utility.*

Definition 2 *A second mover undertakes an action that exhibits* ***positive reciprocity*** *if the chosen action:*

(a) *follows a positive transfer of money by the first mover;*

(b) *gives the first mover a monetary gain; and*

(c) *is undertaken instead of an available alternative action that would produce outcomes preferred by the second mover in the absence of the action by the first mover.*

The concepts of fear and negative reciprocity used in this paper are defined as follows.

Definition 3 *A first mover undertakes an action that exhibits* ***fear*** *if, in two otherwise-identical environments, he:*

(a) *takes money from the second mover when the second mover does not have an opportunity to retaliate; and*

(b) *takes less money or none from the second mover when the second mover does have an opportunity to retaliate.*

Definition 4 *A second mover undertakes an action that exhibits* ***negative reciprocity*** *if the chosen action:*

(a) *follows a non-positive transfer of money by the first mover;*

(b) *reduces the first mover's money payoff; and*

(c) *is an action that would, in the absence of the hurtful action by the first mover, be dispreferred by the second mover to an available alternative action.*

Note that the above definitions of *observable* positive and negative reciprocity incorporate a possible dependence of the *inferred* preferences over outcomes on the process that generates those outcomes and attributions of the intentions of others. And the definitions of *observable* trust and fear incorporate a possible dependence of the *inferred* motivations on the process that generates the outcomes. The triadic experimental design explained in section 2 makes it possible to discriminate between the implications of other-regarding preferences and trust, fear, or reciprocity.

Assumptions

Assume that the other-regarding preferences of an agent can be represented by a utility function, $u(.,.)$ that has the following properties:

1. convexity;
   $\forall P, Q \in \mathcal{R}^2, \forall \lambda \in [0,1]$,
   $$(u(P) \geq c \wedge u(Q) \geq c) \Rightarrow u(\lambda P + (1-\lambda)Q) \geq c. \tag{1}$$

2. egocentricity;
   $$u^1(m,y) > u^1(y,m), \tag{2}$$
   for all $m, y \in \mathcal{R}^+$ s.t. $m > y$;

3. monotonically increasing in both arguments in regions with $m > y$;
   $\forall Z \in \mathcal{Z} = \{Z \in \mathcal{R}^2 \mid Z_2 < Z_1\}$,
   $$\begin{aligned} u^1(Z) &< u^1(Z_1, Z_2 + r) \\ u^1(Z) &< u^1(Z_1 + r, Z_2) \end{aligned} \tag{3}$$
   where $r > 0$.[1]

   Also assume that the (positively-reciprocal) intentions-conditional other-regarding preferences of an agent have the following property.

[1] All existing reciprocity-free models, including Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness Rabin (forthcoming), and Cox, Sadiraj, and Sadiraj (2002), satisfy this assumption.

4. for all actions $A \in R^1_+(I)$ chosen by the first mover, for all actions $P \in R^2(A)$ chosen by the second mover in response[2]

$$P_2 > A_1. \tag{4}$$

Notation

$I = (I, I)$ - the initial allocation of money;

$R^1(I)$ - the budget line of the first mover, that is

$$R^1(I) = R^1_+(I) \cup R^1_-(I) \cup R^1_0(I)$$

where

$$\begin{aligned} R^1_+(I) &= \{A = (I - s, I + 3s) \mid s \in (0, I]\}, \\ R^1_-(I) &= \left\{A = (I - s, I + s) \mid s \in [-\frac{I}{2}, 0)\right\}, \\ R^1_0(I) &= \{I\}; \end{aligned}$$

A - the allocation of money chosen by the first mover in the moonlighting game;

$R^2(A)$ - the budget line of the second mover, that is

$$R^2(A) = R^2_+(A) \cup R^2_-(A) \cup R^2_0(A),$$

where

$$\begin{aligned} R^2_+(A) &= \{P = (A_1 + r, A_2 - r) \mid r \in (0, A_2]\}, \\ R^2_-(A) &= \left\{P = (A_1 + 3r, A_2 + r) \mid r \in [-\frac{A_1}{3}, 0)]\right\}, \\ R^2_0(A) &= \{A\}; \end{aligned}$$

B - the choice of the first mover in the treatment B dictator game;

C - the choice of the second mover in the treatment C dictator game;

[2] Assumption 4 can be interpreted as "positive reciprocity does not dominate egocentricity." This assumption is satisfied by 100% of our data.

P - an allocation of money chosen by the second mover in the moonlighting game;

T - the intersection of the budget line of the first mover with the vertical axis;

Some Results

Lemma 5 *1.* $\forall Z \in \{(m,y) \in \mathcal{R}^2_+ \mid m < y\}, \forall \varepsilon \in (0, y-m)$

$$u^1(m,y) < u^1(m+\varepsilon, y-\varepsilon);$$

2. $\forall A \in R^1(I)$,

(a) $A \notin R^1_+(I) \Rightarrow P \notin R^2_+(A)$,
(b) $A \in R^1_+(I) \Rightarrow P \notin R^2_-(A)$;

3. B does not exhibit either trust or fear;

4. C does not exhibit any type of reciprocity.

Proof.

1. Let some $Z \in \{(m,y) \in \mathcal{R}^2_+ \mid m < y\}$ be given. Take $Z^* = (y,m)$. From the egocentricity property (2) $u^1(Z) < u^1(Z^*)$. Since $(m+\varepsilon, y-\varepsilon)$ is a convex combination of Z and Z^*, for all $\varepsilon \in (0, y-m)$ the convexity property (1) completes the proof.

2.

   (a) Let the allocation of money chosen by the first mover in the moonlighting game be some $A \notin R^1_+(I)$,that is $A = (I-s, I+s)$ for some $s \in [-I/2, 0]$. Suppose that $P \in R^2_+(A)$. Note that agent 1 has not been transfering a positive amount of money to agent 2 (since $s < 0$) and therefore an action undertaken by the second mover cannot be characterized by the positive reciprocity since it violates part (a) of definition 2. Since P gives a monetary gain to the first mover (part (b) of definition 4 is violated) the second mover's behavior cannot be characterized by negative reciprocity either. Therefore, the second mover giving the first

mover a monetary gain could be motivated only by other regarding intentions-unconditional preferences and hence the following must be true

$$u^2(P) > u^2(A) \tag{5}$$

On the other hand, applying Lemma 5.1 for the second mover, one has

$$u^2(P) < u^2(A) \tag{6}$$

which contradicts (5). Thus, $P \notin R^2_+(A)$.[3]

(b) Let $A \in R^1_+(I)$. That implies that the second mover is given money and therefore his choice cannot be motivated by negative reciprocity. Note that first, the positive direct reciprocity motivation would induce a choice from $R_+(A)$ since the part (b) of the definition 2 of positive reciprocity requires agent 1 receive a monetary gain, i.e. $r > 0$; and second, property (3) of the intentions-unconditional other-regarding preferences implies

$$u^2(A) > u^2(P), \text{ for all } P \in R_-(A)$$

and therefore $P \notin R_-(A)$.[4]

3. Suppose B is chosen by the first mover in the moonlighting game. There can be only two cases:

Case 1 $B \notin R^1_+(I)$.

Part (a) of the definition 1 of trust is violated since the sum of money payoffs $(2I)$ does not increase. Therefore B cannot be a trusting action.

Furthermore, if $B \in R^1_-(I)$ then B cannot exhibit fear since part (b) of the definition 3 of fear is not satisfied. If $B \in R^1_0(I)$ then

[3] This proof applies to treatment C and implies that the following hypothesis should be true (for treatment C):

Ho: if $A_1 \geq A_2$ then $r^c(A) \leq 0$

88% of the choices made in Treatment C by subjects faced with initial allocation of money such that $A_1 \geq A_2$ confirm the above null hypothesis.

[4] This proof applies to treatment C and implies that the following hypothesis should be true (for treatment C):

Ho: if $A_1 < A_2$ then $r^c(A) \geq 0$

100% of the choices made in Treatment C by subjects faced with initial allocation of money such that $A_1 < A_2$ confirm the above null hypothesis.

B cannot exhibit fear since part (b) of the definition 3 of fear is not satisfied;

Case 2 $B \in R^1_+(I)$

Part (b) of the definition 1 of trust is violated as the following shows. First, note that property (4) implies that $P_2 > B_1$ for any chosen action $P \in R^2(B)$ by the second mover in response to the action B chosen by the first mover. Second, we show that

$$u^1(P) \geq u^1(B), \text{ for all } P \in R^2_+(B) \cup \{B\}, P_2 > B_1 \qquad (7)$$

Indeed, let $P \in R^2_+(B) \cup \{B\}$, $P_2 > B_1$ be given. If $P = B$ then the equality holds. Otherwise $P = (I - s^b + r, I + 3s^b - r)$ for some $r \in (0, 4s^b)$[5] where $s^b > 0$. Taking $m = I - s^b, y = I + 3s^b$, $\varepsilon = r \in (0, y - m)$, and applying 5.1 one has

$$u^1(B) < u^1(P);$$

Third, the monotonicity property (3) in the region with $m > y$ implies the second mover's choice be not from $R^2_-(B)$. Noting that the second mover choice cannot be motivated by negative reciprocity because $B \in R^1_+(I)$ (i.e. the first mover has been transfering positive amount of money to the second mover, that is part (a) of definition 4 is violated), one completes the proof.

B does not exhibit fear since part (a) of the definition 4 of fear is not satisfied.

4. $P = C$ violates both definitions $2(c)$ and $4(c)$.

■

Proposition 6 *Let a first mover undertake an action A and a second mover undertake an action P in response.*

1. *(a) $A \in R^1_+(I)$ is a necessary condition for exhibiting trust but not a sufficient one;*

   (b) $A \in R^1_+(I)$ and $A_2 > B_2$ are sufficient conditions for exhibiting trust;

2. *A exhibits fear if and only if $A \notin R^1_+(I)$ and $A_2 > B_2$;*

[5] $P_2 > A_1$ is equivalent to $I + 3s^b - r > I - s^b$ and therefore $r < 4s^b$.

3. P exhibits positive reciprocity if and only if $P \in R^2_+(A)$ and $P_2 < C_2$;

4. P exhibits negative reciprocity if and only if $P \in R^2_-(A)$ and $P_2 < C_2$.

Proof.

1.

   (a) Let A exhibit trust. We show that $A \in R^1_+(I)$.
   Suppose that $A \notin R^1_+(I)$. Since $A \in R^1(I)$ one has $A \in R^1_-(I) \cup R^1_0(I)$, and therefore there is no monetary gain created by the first mover, which contradicts part (a) of the definition 1 of trust.

   Next, we show that the condition is not sufficient. Indeed let the action chosen by the first mover in the dictator game be $B \in R^1_+(I)$. Then $A = B \in R^1_+(I)$ and from Lemma 5.3 A does not exhibit trust.

   (b) Let $A \in R^1_+(I)$, and $A_2 > B_2$. $A \in R^1_+(I)$ implies that $A = (I - s, I + 3s)$ for some $s \in (0, I]$. Hence a monetary gain of $2s$ is created that could be shared, so part (a) of the definition 1 of trust is satisfied. $A_2 > B_2$ implies that $A \neq B$. B is a tangent point; hence property (1) implies that $u^1(A) < u^1(B)$. Since A can be chosen by the second mover, and since A satisfies part (b) of the definition 1 of trust, A is a trusting action.

2. Let A exhibit fear. Note that $A \in R^1_+(I)$ violates part (b) of the definition 3 of fear. Furthermore, part (a) of the definition 3 of fear implies $B \in R^1_-(I)$. Now we show that $A \in R^1_-(I)$,and $A_2 \leq B_2$ cannot be true. $A_2 = B_2$ cannot be true as Lemma 5.3 shows. Suppose $A \in R^1_-(I)$ and $A_2 < B_2$. Note that A satisfies (i) $u^1(A) < u^1(B)$ (B is the tangent point); and $(ii) u^2(A) < u^2(C)$ (apply Lemma 5.1). Thus, C improves the utilities of both agents and therefore dominates A. Concluding, $A_2 > B_2$.

   Now let $A \notin R^1_+(I)$ and $A_2 > B_2$. $A_2 > B_2$ and $A \notin R^1_+(I)$ implies $B \in R^1_-(I)$ and hence part (a) of the definition 3 of fear is satisfied. $A_2 > B_2$ and $A \notin R^1_+(I)$ implies part (b) of the definition 3 of fear is satisfied.

3. Let P be a positively reciprocal action. That implies $P \in R^2_+(A)$ since $P \in R^2_-(A) \cup R^2_0(A)$ would violate part (b) of the definition 2 of positive reciprocity. Let the action chosen by the second mover in

the dictator game be $C \in R^2(A)$. From Lemma 5.4 one has $C \neq P$. From part (a) of the definition 2 of positive reciprocity agent 2 is transferred a positive amount of money, and therefore he is allocated the higher monetary payoff. Hence from property (3), $C \notin R^2_-(A)$. Thus $C \in R^1_+(A) \cup R^1_0(A)$. Let $C \in R^1_0(A)$. $P_2 \leq A_2 = C_2$ and $P \neq C$ imply $P_2 < C_2$. Now let $C_2 \in R^1_+(A)$. First, note that $P \in R^2_+(A)$, $P_2 > C_2$ can never be chosen by the second mover as the following shows. For any given $P \in R^2_+(A), P_2 > C_2$ one has (i) $u^2(P) < u^2(C)$ (C is the tangent point), and $(ii) u^1(P) < u^1(C)$ (note that $C_2 > C_1$ from Lemma 5.1 and apply the same Lemma for agent 1). Thus, C improves the utilities of both agents and therefore dominates P. Hence, $P_2 < C_2$.

Now let $P \in R^2_+(A)$ and $P_2 < C_2$. From Lemma 5.2a, $P \in R^2_+(A)$ implies $A \in R^1_+(I)$ and therefore part (a) of the definition of 2 of positive reciprocity is satisfied. $P \in R^2_+(A)$ implies part (b) of the definition 2 of positive reciprocity is satisfied. $P_2 < C_2$ implies $P \neq C$ and thus part (c) of the definition 2 of positive reciprocity is satisfied.

4. Let $P \in R^2_-(A)$ and $P_2 < C_2$. From Lemma 5.2b, $P \in R^2_-(A)$ implies $A \notin R^1_+(I)$ and therefore part (a) of the definition of 4 of negative reciprocity is satisfied.$P \in R^2_-(A)$ implies part (b) of the definition 4 of negative reciprocity is satisfied. $P_2 < C_2$ implies $P \neq C$ and thus part (c) of the definition 4 of negative reciprocity is satisfied.

   Now let P be a negatively reciprocal action. That implies $P \in R^2_-(A)$ since $P \in R^2_+(A) \cup R^2_0(A)$ would contradict part (b) of the definition 4 of negative reciprocity. Let the action chosen by the second mover in the dictator game be C. Lemma 5.4 implies $C \neq P$. From Lemma 5.2b, $P \in R^2_-(A) \Rightarrow A \notin R^1_+(I)$. $A \notin R^1_+(I)$ implies $A_2 \leq A_1$ and hence applying Lemma 5.1 $C \notin R^2_+(A)$. Thus, $C \in R^2_-(A) \cup R^2_0(A)$. If $C \in R^2_0(A)$ then $P_2 \leq A_2 = C_2$ and $P \neq C$ implies $P_2 < C_2$. If $C \in R^2_-(A)$, then first, we show that $P \in R^2_-(A)$, and $P_2 > C_2$ can never be chosen by the second mover. For this, first, note that from property (3) in a region where $m > y$ for the second mover one derives that $C_2 \leq C_1$, and second note that since (i) $u^2(P) < u^2(C)$ (C is the tangent point) and (ii) $u^1(P) > u^1(C)$ (from property (3) in the region where $m > y$ for the first mover), C improves the utility of agent 2 and hurts more agent 1, therefore dominates P. Hence, $P_2 < C_2$.

■