Implications of Trust, Fear, and Reciprocity for Modeling Economic Behavior

By James C. Cox, Klarita Sadiraj, and Vjollca Sadiraj

This paper uses a three-games (or triadic) experimental design to discriminate between actions motivated by unconditional preferences over the distribution of material outcomes and actions motivated by attributions of the intentions of others and beliefs about their behavior. 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 in positive reciprocity or fear of negative reciprocity, in addition to other-regarding preferences. Second movers can be motivated by other-regarding 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 is characterized by trust and insignificant fear and that second movers' behavior is characterized by positive reciprocity and insignificant negative reciprocity.

Keywords: experiments, game theory, intentions, beliefs
JEL Classification: C70, C91, D63, D64

1. Introduction

Applications of game theory in economics 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 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 or by beliefs about another's reciprocity. But actions that are inconsistent with self-regarding preferences can also be motivated by agents' altruistic or inequality-averse other-regarding preferences that do not reflect beliefs about others' future actions or attributions of the intentions revealed by their past actions.[1] In this paper, the term "unconditional preferences" refers to beliefs- and intentions-

unconditional preferences over outcomes such as the altruistic preferences that are revealed by charitable contributions to organizations without political agendas. In contrast, "conditional preferences" will be used to refer to beliefs- and/or intentions-conditional preferences over outcomes such as the negatively-reciprocal preferences that are revealed when, in response to a hurtful action, an agent reduces his own material payoff in order to harm the person responsible for the original hurtful action.

The distinction between actions motivated by beliefs- and intentions-unconditional preferences over outcomes and actions motivated by beliefs or attributions of intentions is essential to empirical guidance for theory development because modeling beliefs or intentions is quite different from modeling unconditional preferences over outcomes. The importance of the distinction between intentions-conditional and intentions-unconditional preferences over outcomes for modeling behavior in mini-ultimatum, Stackelberg duopoly, and moonlighting games is shown by the model developed in Cox, Friedman, and Gjerstad (2004).

In order to obtain data that can guide development of economic models that are consistent with behavior, 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 unconditional preferences over outcomes and actions dependent on beliefs about others and/or attributions of their 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 beliefs about the other subject's actions because the paired subject has no choice to make.

Results from earlier experiments with the moonlighting game by Abbink, Irlenbusch, and Renner 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/or inequality-averse preferences over

outcomes that are not conditional on attributions of others' intentions. A more elaborate experimental design can provide essential empirical support for alternative modeling strategies. The central idea of our work is to construct control environments that reveal whether behavior in a central game of interest (in the present case, the moonlighting game) can be represented with a model of unconditional other-regarding preferences or, instead, requires construction of a more complicated model that incorporates agents' attribution of intentions and/or beliefs about others.

There are a few other studies that use control treatments for intentions. Blount (1995), Charness (forthcoming) and Offerman (2002) use random selection of first moves as intentions-control treatments in various games. Bolton, Brandts, and Ockenfels (1998) use an intentions-control treatment in which the row player is given the task of "choosing" between two identical rows of monetary payoffs in simple dilemma games. The present paper differs from the above papers in that it uses controls for both second movers' attributions of intentions and first movers' beliefs about the second movers' actions. Cox (2002, 2004) uses both intentions- and beliefs-control treatments in experiments with the investment game in which there can be positive reciprocity and trust in positive reciprocity. The present paper generalizes this approach by using the moonlighting game in which there can be both positive and negative reciprocity, trust in positive reciprocity, and fear of negative reciprocity.

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.[2] At the time the experiment sessions were run, 1 euro was worth a little less than 1 dollar.[3]

2.1. The Three Games

Treatment A is the moonlighting game. Each individual is either a first-mover or a second-mover. Each second-mover is credited with a 10 euro endowment. Each first-mover is credited with a 10 euro endowment and given the task of deciding whether she wants to give to a paired second-mover 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 second-mover 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 three euro that the second mover takes from the paired first mover costs the second mover one 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.a shows representative choices open to the subjects in treatment A. The piece-wise-linear solid line passing through points T, B, and I 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, she would change the two subjects' endowments from (10,10) to (3,31), indicated by point T in Figure 1.a. This choice by the first mover would determine the second mover's "budget line" shown by the piece-wise-linear dashed line with a kink at point T in Figure 1.a. The slope of the second mover's "budget line" is $+1/3$ below the first mover's budget line and -1 above it. If instead, the first mover were to take 4 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 F in Figure 1.b.

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 has the same budget line as in treatment A, shown by the piece-wise linear solid line passing through points T, B and I in Figure 1.a. The first mover chooses an integer amount to give or take from the "second mover." 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, a "first mover" does not have a decision to make. A "second mover" is given one of the "budget lines" determined by a first mover's decision in treatment A. The "second mover" then determines both subjects' money payoffs by choosing an integer amount on her "budget line" to give to or take from the paired "first mover."

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 sealed 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 rooms by a floor to ceiling partition. One room was accessed through the door marked # and the other room was accessed through the door marked ∗. The windows between the experimenters' control room and the laboratory were covered by blinds. Thus, subjects in the two rooms 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. Each subject also had a printed version of the instructions available for review during the decision-making part of the experiment. The instructions referred to the subjects only as being type X or type 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 online on an author's webpage (http://uaeller.eller.arizona.edu/~jcox/index.html).

The decision-making part of the experiment proceeded as follows. First, the monitor computer randomly determined which room, # or * was the room with type X subjects and which was the room with type Y subjects. The pairing of type X and type 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 type X person had a beginning account balance of $10 + A_j$ and the type Y person had a beginning account balance of $10 + B_j$. The amounts, A_j and B_j were determined by type X individuals' 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 type 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.[4] 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 a "level 1 attribution" involving perception 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. The decision to use an across-subjects protocol was made after careful consideration of the respective advantages and disadvantages of within-subjects and across-subjects protocols in the context of intentions- and beliefs-control experiments. The central idea of such experiments is to construct treatments that *reveal* whether behavior in a central game of interest (in the present case, the moonlighting game) can be represented with a model of unconditional other-regarding preferences or, instead, requires construction of a more complicated model that incorporates agents' attribution of intentions revealed by others' past actions and/or beliefs about others' future actions. In order to pose these empirical questions, the dictator control treatments provide subjects with the same (own income, other's income) feasible choice sets as does the central game but remove the decision opportunity of the paired subject, and thereby remove the possible effects of beliefs and intentions attribution on behavior. Ideally, one would like to be able to implement the control treatments with a within-subjects protocol in order to "decompose" the motivations of specific subjects. But available data (Cox, 2003, Fehr and Schmidt, 2003) support the view that this approach is not feasible because giving subjects more than one decision alters their behavior.[5] It appears to be impossible to obtain the requisite "other-things-equal-across-treatments" property of intentions- and beliefs-

control treatments with a within-subjects protocol. Therefore, we use an across-subjects protocol that decomposes the motivations of subject samples from a single population.

3. Identifying the Effects of Beliefs and Intentions on Behavior

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. Suppose that an agent's unconditional preferences can be represented by a utility function. In the special case of two-agents, the agent has "self-regarding" preferences if his utility function is an increasing function of his own monetary payoff and is a constant function of the other's money payoff. The agent has "other-regarding" preferences if her utility function, is *not* a constant function of the other's money payoff. If the agent's utility function is everywhere positively monotonic in the other's money payoff then preferences are said to be altruistic (see Cox, Sadiraj, and Sadiraj, 2002 for an altruistic model with the "ego-centricity" property). The preferences are inequality-averse (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) if the agent's utility function is positively monotonic in the other's money payoff when the other person is allocated less money and negatively monotonic in the other's money income otherwise. It is important to distinguish actions motivated by (unconditional) other-regarding preferences from actions motivated by conditional preferences such as perceived intentions and/or beliefs such as trust, fear, or reciprocity.

3.1. Identifying Trusting Behavior

A trusting action is an action that generates a profit that can be shared and requires a belief by the first mover that the second mover will not defect and keep too much of the profit. If a first mover has either inequality-averse or purely self-regarding preferences then she will have indifference curves in Figure 1.a that either have positive slopes above the 45-degree line or are vertical

straight lines. In that case, the act of sending any positive amount implies trust because it has created a profit that could be shared and it reveals a first-mover belief that the second mover will return a sufficient amount of money to leave the first mover be at a higher utility level that at the endowment point, *I*. But a first mover may have altruistic preferences and, hence, indifference curves with negative slopes above the 45-degree line. Since, in the moonlighting game any positive amount sent by the first mover increases the account balance of the second mover, the slope of the first mover's budget line above the 45-degree line is negative. Therefore, a first mover with altruistic preferences might prefer to give the second mover some money regardless of how much, if any, the second mover might return. 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, as follows.

Suppose that the amount of money that the first mover sends to the second mover in treatment A, s_a is positive, which creates a profit (of $2s_a$) that can be shared. If, in addition, s_a is 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 unconditional preferences. Thus, if the amount of money sent in Treatment A, s_a is such that

(1) $\quad s_a > 0$ and $s_a > s_b$

then we can conclude that the first mover's choice has been motivated by a belief that the second mover will return enough money to leave the first mover with utility no lower than it would be if he were to set $s_a = s_b$ and the second mover were to return 0.

Figure 1.a illustrates a trusting action by a first mover with altruistic preferences. An example of an altruistic choice, $s_b > 0$ is shown by the tangency of the first mover's (unconditional) indifference curve with his "budget line" at the point B in Figure 1.a. An

example of a trusting choice $s_a (> s_b > 0)$ is shown by the point T in Figure 1.a. This choice gives the second mover the piece-wise-linear dashed budget line with a kink at point T. If the second mover chooses any amount to return that would result in some money allocation 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 1.a) and the second mover had returned, for example, any amount of money from 0 to $2 s_b$.

3.2. Identifying Positively-Reciprocal Behavior

Next consider the question of identifying positively-reciprocal behavior. The preferences over payoff 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 sent 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. 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.

Note that a second mover with either altruistic or inequality-averse preferences may give money to a first mover who has a lower money endowment than the second mover because other-regarding preferences (both altruistic and inequality-averse types) are positively monotonic in the other's money payoff when the other is allocated the smaller amount of money. Thus the mere fact that the second mover returns a positive amount of money to a first mover who gave money to the second 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.

Consider a "second mover" in treatment C who is given an endowment that is larger than 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, s_a (but the subjects do not know this). In treatment C, a "second mover" chooses an amount to return, r_c from a feasible set of integers on the piece-wise-linear "budget line" determined by a first mover's choice of s_a in treatment A. An example of a possible choice, r_c is shown by the tangency of the "second mover's" (unconditional) indifference curve with her "budget line" at point C in Figure 1.a.

Suppose that the second mover returns to the first mover in the moonlighting game, treatment A, a positive amount of money. 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 1.a). But if one observes that $r_a > r_c$ then he can conclude that the second mover was motivated by positive reciprocity because the amount of money returned is too large to be fully accounted for by unconditional other-regarding preferences. In Figure 1.a, 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. Thus, it follows from the generosity property of positive reciprocity, the optimality of r_c and strict-quasiconcavity of the unconditional other-regarding preferences that all actions of the second mover, r_a such that

(2) $\quad r_a > 0$ and $r_a > r_c$

exhibit positive reciprocity.

3.3. Identifying Fearful Behavior

We next turn our attention to fear of negatively-reciprocal (or punishing) behavior. A fearful action requires a belief by the first mover that the second mover has inequality-averse preferences

and/or that he will retaliate and punish the first mover's decision to take money from the second mover. With respect to unconditional preferences, it might be an optimal action for a first mover to take money from the second mover. The central question is whether or not the first movers in the moonlighting game who do not trust take less money, if any, from the second movers when the second movers can make a decision, than they would in the absence of the second mover's opportunity to retaliate. The treatment B dictator game together with the treatment A moonlighting game permit one to identify fearful actions, as follows.

Suppose that the first mover takes money from the second mover in treatment A. If, in addition, $|s_a|$ is smaller than the amount taken in treatment B, $|s_b|$ then one can conclude that the first mover's behavior is motivated by fear because the amount taken from the second mover in treatment A is too small to be fully explained by unconditional preferences. Thus, it follows from strict quasiconcavity of unconditional preferences, optimality of s_b, and the greed-restraining property of fear that

(3) $\quad s_b < s_a \leq 0$

is a sufficient condition for fearful behavior. Therefore, one must incorporate fearful beliefs into the model in order to explain the first mover's behavior if the data satisfy the inequalities in statement (3).

Figure 1.b illustrates behavior of a fearful first mover. In that figure the first mover's choice in treatment B is at the tangency point B whereas point F is the first's mover choice in treatment A. Point F is on a lower (unconditional) indifference curve for the first mover than is point B. The choice of F in the moonlighting game is evidence that the first mover is afraid to choose his most preferred outcome point B in the moonlighting game.

3.4 Identifying Negatively-Reciprocal Behavior

Finally, consider the question of identifying negatively-reciprocal behavior. The preferences over

payoff pairs can be conditional on a social norm for reciprocating an unfriendly action of another. For example, if the second mover knows that the first mover in the moonlighting game took money from the second mover, the second mover may be motivated by a social norm for negative reciprocity to punish that greedy behavior. The empirical question is whether or not second movers in the moonlighting game choose harsher actions, after the first mover has taken money from them, than they would in the absence of the first mover's action but the presence of the same money allocation.

Note that a second mover with unconditional inequality-averse preferences may prefer a costly action which decreases the money payoff of the other player who has a larger money endowment than the second mover in order to reduce the difference between two endowments. Thus, the mere fact that a second mover takes money from a first mover who previously took money from the second mover is not evidence of negative reciprocity. But the treatment C dictator game, together with the treatment A moonlighting game, permit one to identify negatively-reciprocal actions, as follows.

Consider a "second mover" in treatment C who is given an endowment that is smaller than the endowment of the paired subject. Recall that the endowments of a pair of subjects in treatment C are determined by a (distinct) first mover's decision in treatment A, s_a (but the subjects do not know this). In treatment C, a "second mover" chooses an amount to return, r_c from a feasible set of integers on the piece-wise-linear "budget line" determined by a first mover's choice of s_a in treatment A. Suppose that the second mover makes a costly decision, $r_a < 0$ which reduces the first mover's money payoff in the moonlighting game, treatment A by $3|r_a|$. This, in itself, does not support a conclusion that the second mover was motivated by negative reciprocity because the assumed choice could have been motivated by unconditional inequality-averse preferences, as shown by the tangency of the upward-sloping part of the

(unconditional) indifference curve at point C in Figure 1.b. However, if in addition one observes that $r_a < r_c$ then he can conclude that the second mover was motivated by negative reciprocity because the hostile action is too costly to be fully accounted for by unconditional (inequality-averse) preferences. In Figure 1.b, any location of r_a on the second mover's "budget line" that is downwards and to the left of point C would exhibit negative reciprocity.

Thus, the optimality of r_c, strict quasiconcavity of unconditional preferences, and the punishing property of negative reciprocity imply that

(4) $r_a < 0$ and $r_a < r_c$

is a sufficient condition for exhibiting negative reciprocity.

4. Subjects' Behavior in the Experiment

We first describe the subjects' behavior in the moonlighting game and, subsequently, present tests of hypotheses using data from all three games.

4.1. First- and Second-Mover Choices in the Moonlighting Game

The subjects' behavior in treatment A, the moonlighting game is presented in Figure 2. 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. The behavior of these 12 subjects is consistent with the "economic man" model of unconditional self-regarding preferences. One subject took one euro, three subjects "sent" zero, and 14 subjects gave positive amounts to the second mover; five subjects gave the maximum of 30 euro. The behavior of these 18 subjects is inconsistent with the model of unconditional self-regarding preferences.

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 loss of 0.33 euro or an average profit of 0.15 euro after the second movers made their decisions, depending on whether one anomalous observation is excluded or included.[6] In any case, one concludes that, on average, giving was more profitable than taking by first movers because 1.93 is larger than both -0.33 and 0.15.

Next consider the behavior of second movers. Note that 13 of the 30 second movers neither gave nor took money from first movers. The behavior of these 13 subjects is consistent with the model of unconditional self-regarding preferences. 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. The behavior of these 17 subjects is inconsistent with the model of unconditional self-regarding preferences.

We have observed that the behavior of 35 out of the 60 subjects in the moonlighting game cannot be rationalized by the unconditional self-regarding preferences model. Therefore, some type of other-regarding preferences is needed to rationalize the behavior of more than half the subjects. The triadic design produces data that tell us whether the subject pool consists of subjects whose behavior can be rationalized by models of unconditional other-regarding preferences or, instead, includes subjects whose preferences over outcomes are conditional on beliefs about others' future actions or attributions of the intentions underlying their past actions.

4.2 Tests for Trust, Fear and Reciprocity

Consider the behavior of first movers. Figure 3 presents the amounts sent in treatments A and B. First movers' behavior appears to be more generous in treatment A than in treatment B. Altruism could motivate sending positive amounts in either treatment A or B. In contrast, first movers' trust that second movers will return a sufficiently large amount of money 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.

A 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 3, 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. A test for the significance of trusting behavior can be constructed as follows. Referring to the inequalities in statement (1) (see section 3.1), an action of sending an amount of money, s_a exhibits trust if $s_a > x \equiv \max\{s_b, 0\}$. Hence, the null hypothesis we test is $s_a = x,$ and the alternative is $s_a > x.$ The null hypothesis is rejected at 5% significance level according to the one-tailed Smirnov test and the means test (see the third row of Table 1). We conclude that first movers' behavior in the moonlighting game is characterized by trust. Therefore, this behavior cannot be rationalized by models of unconditional other-regarding preferences because the revealed preferences over outcomes are conditional on beliefs about second movers' future actions.

Next consider the question of whether the first movers' behavior in treatment A is characterized by fear of negative reciprocity. 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 unconditional inequality-averse preferences, or he may even send money in treatment A if he trusts in positive reciprocity and/or altruistic preferences.

Referring to the inequalities in statement (3) (see section 3.3), the null hypothesis we test is $s_a = s_b$ and the alternative is $s_a > s_b$ for $s_a \in \{s \in A \mid s \leq 0\}$ and $s_b \in \{s \in B \mid s \leq 0\}$. The null hypothesis is rejected at the 5% significance level by the means test but not by the Smirnov test (see the fourth row of Table 1). This provides additional (weak) support for the conclusion that first movers' preferences are conditional on beliefs about second movers' future actions.

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 direct reciprocity in treatment A but not in treatment C. Whether or not the behavior of second movers *is* characterized by direct reciprocity is revealed by comparing responses in treatments A and C.

Figures 2 and 4 show how second movers responded to amounts they received in treatments A and C. Figure 5 presents a direct comparison of amounts returned in treatments A and C, and also shows the amounts sent in treatment A. 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 second movers responded by returning positive amounts to first movers and three second movers kept all of the money. In contrast, in treatment C four "second movers" gave positive amounts to first movers and 10 "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 the second movers who received 30 euro returned positive amounts, with the amounts returned varying from a low of 10 euro to a

high of 20 euro. Finally, note that the fifth row of Table 1 reports tests comparing amounts returned in treatments A and C by second movers who received positive amounts. Both 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. Therefore, second movers' preferences over outcomes are conditional on the revealed intentions underlying first movers' actions.

Next consider responses by the 13 second movers who "received" negative amounts in both treatments. In treatment A, 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 1 reports tests comparing amounts returned in treatments A and C by second movers who received negative amounts. Both tests do not reject the null hypothesis of an absence of negative reciprocity at 5% significance level.[7] We conclude that second movers' behavior is not characterized by significant negative reciprocity.

5. Concluding Remarks

This paper reports an experiment with a game triad that includes the moonlighting game. Results from the experiment support the conclusion that the first movers in the moonlighting game were motivated by trust that the second movers 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 the first movers is weakly characterized by fear. The behavior of second movers who had money taken from them by first movers was not characterized by significant negative reciprocity.

We conclude that behavior in games involving salient fairness considerations is rich, and that neither the "economic man" model of purely self-regarding preferences nor a model of unconditional other-regarding preferences is sufficiently rich to explain behavior in such games. Models that include other-regarding preferences that are conditional on beliefs about others' actions and perceptions of their intentions are needed to explain fairness-game behavior. 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

1. Models of unconditional inequality-averse preferences are presented in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). Models of unconditional altruistic other-regarding preferences are presented in Andreoni and Miller (2002), Charness and Rabin (2002), and Cox, Sadiraj, and Sadiraj (2002).
2. 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.
3. 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 earlier experiments (for example, Berg, et al.,1995 and Cox, 2002, 2004).
4. See Dufwenberg, Gneezy, Güth, and van Damme (2001) for tests of both direct and indirect reciprocity in the context of the investment game.
5. For example, telling subjects there will be another decision task after a dictator game significantly shifts their behavior towards greater generosity, even in an experiment in which there is anonymity (because of double-blind payoffs) and random selection of one task for payoff (Cox, 2003). In another example, prior participation in an ultimatum game shifts behavior in single-decision-maker distributional games towards more egalitarian outcomes (Fehr and Schmidt, 2003).

6. Observation 12 in Figure 2 is anomalous because after the first mover took the maximum of five euro from the second mover, the second mover responded by giving the first mover one euro.
7. The same conclusion follows if the tests use the truncated data from which the anomalous observation number 12 in Figure 2 and observations numbered 11 and 12 in Figure 4 are deleted. In that case the means test statistic is –1.32 and the Smirnov test statistic is 0.33.

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 John Miller, "Giving According to GARP: An Experimental Test of the Rationality of Altruism," *Econometrica,* v. 70, no. 2, March, 2002, pp. 737-53.

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

Charness, Gary, "Attribution and Reciprocity in an Experimental Labor Market." *Journal of Labor Economics*, forthcoming.

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

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, 2002.

Cox, James C., "Trust and Reciprocity: Implications of Game Contexts and Social Contexts," Discussion Paper, University of Arizona, 2000; revised 2003.

Cox, James C., "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 46, 2004, pp. 260-281.

Cox, James C., Daniel Friedman, and Steven Gjerstad, "A Tractable Model of Reciprocity and Fairness," Discussion Paper, University of Arizona, 2004; presented at the June 2004 meetings of the Economic Science Association.

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.

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

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.

Fehr, Ernst and Klaus M. Schmidt, "The Role of Equality, Efficiency, and Rawlsian Motives in Social Preferences: A Reply to Engelman and Strobel," Discussion Paper, University of Zürich, 2003.

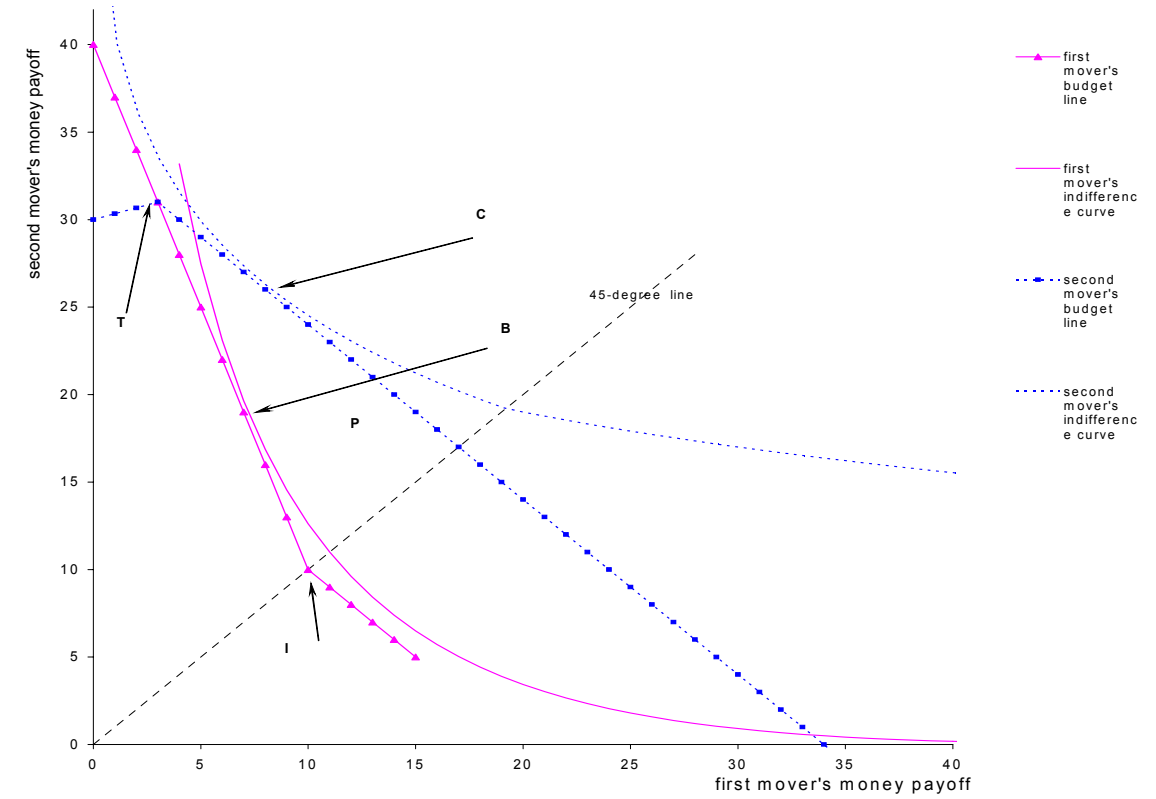
Offerman, Theo, "Hurting Hurts More than Helping Helps: The Role of the Self-Serving Bias." *European Economic Review, 46,* 2002, pp.1423-37.

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

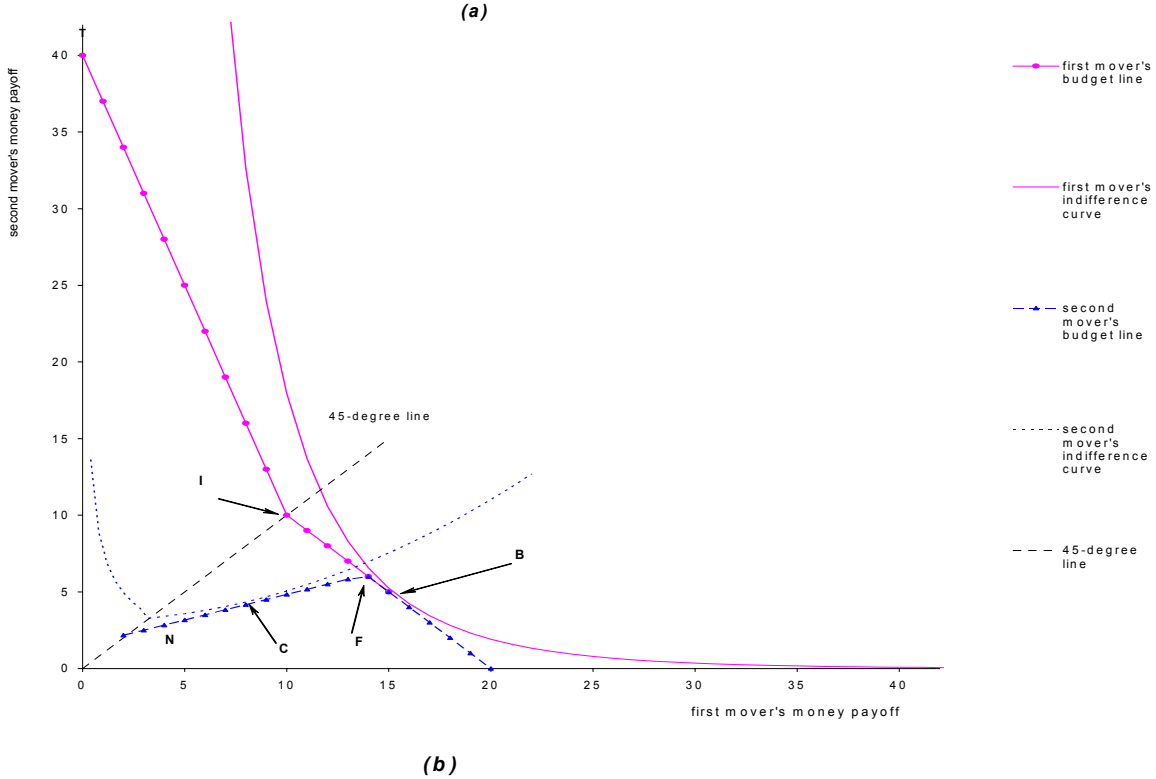
Table 1. Tests for Trust, Fear, and Reciprocity

Data	Return Mean			Means Test	Smirnov 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 [2.91] {30}	…	…
Tr. A Send vs. Tr. B Send (trust)	…			1.78*	-0.396*
Tr. A Send vs. Tr. B Send (fear)	…			2.22*	-0.25
Tr. A Return vs. Tr. C Return (positive reciprocity)	…			4.10*	-0.71*
Tr. A Return vs. Tr. C Return (negative reciprocity)	…			-1.45	0.31

Standard deviations in brackets.
Number of subjects in braces.
* significant at 5%.



(a)



(b)

Figure 1. Illustration of Representative Budget Lines of First and Second Movers, Trust and Positive Reciprocity *(a)* and Fear and Negative Reciprocity *(b)*.

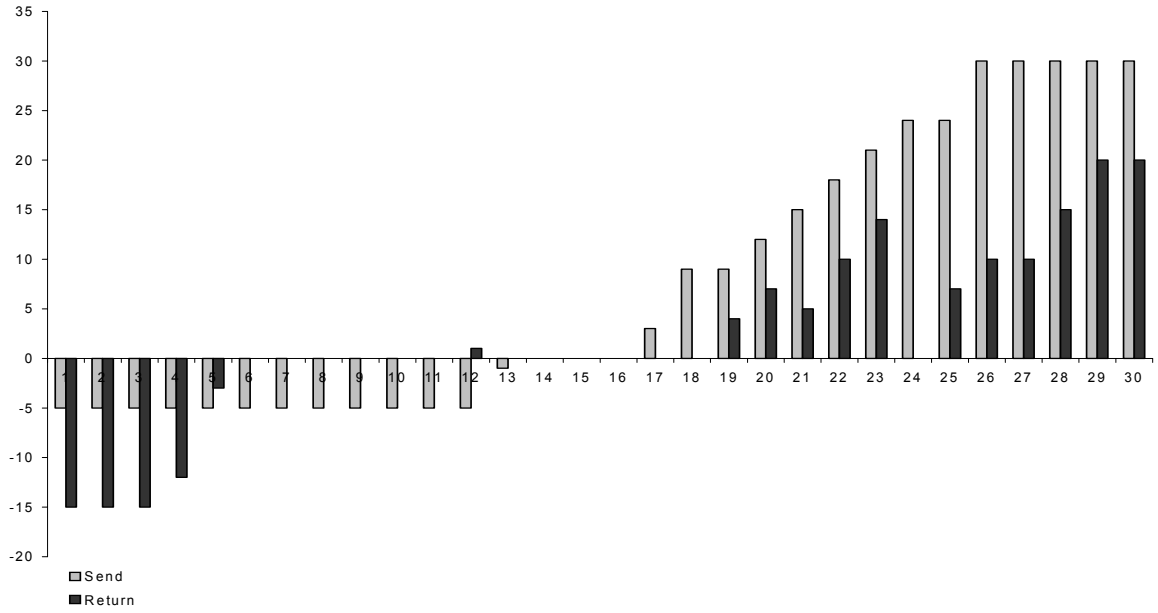


Figure 2. Money Sent and Returned in Treatment A

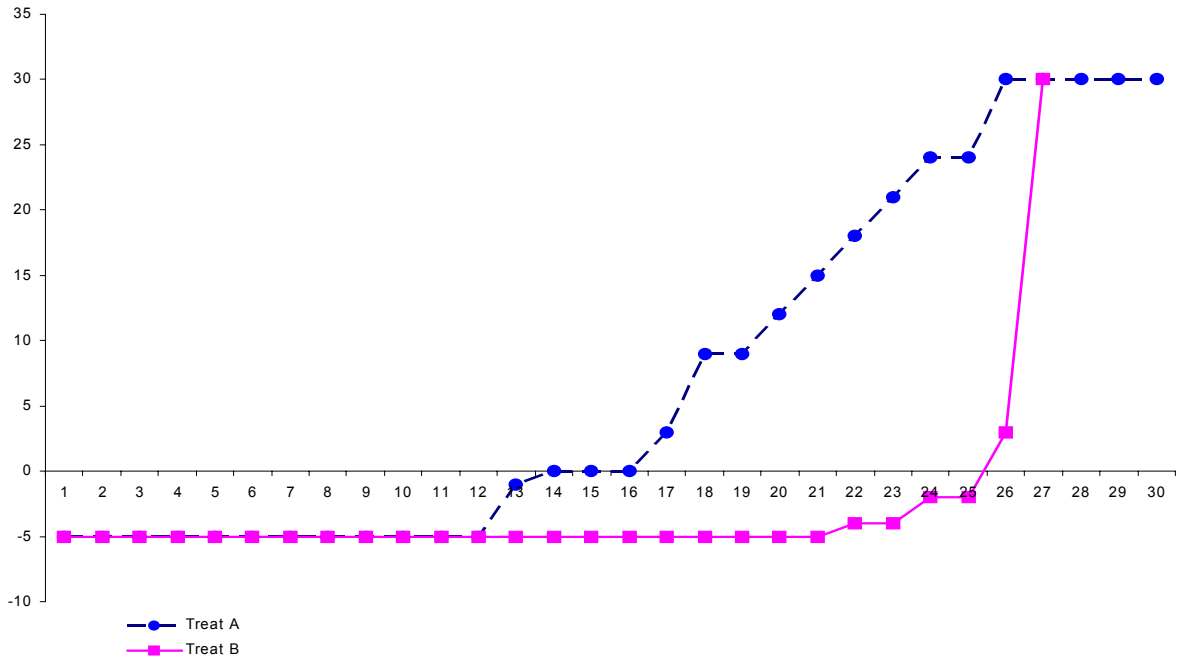


Figure 3. Money Sent in Treatments A and B

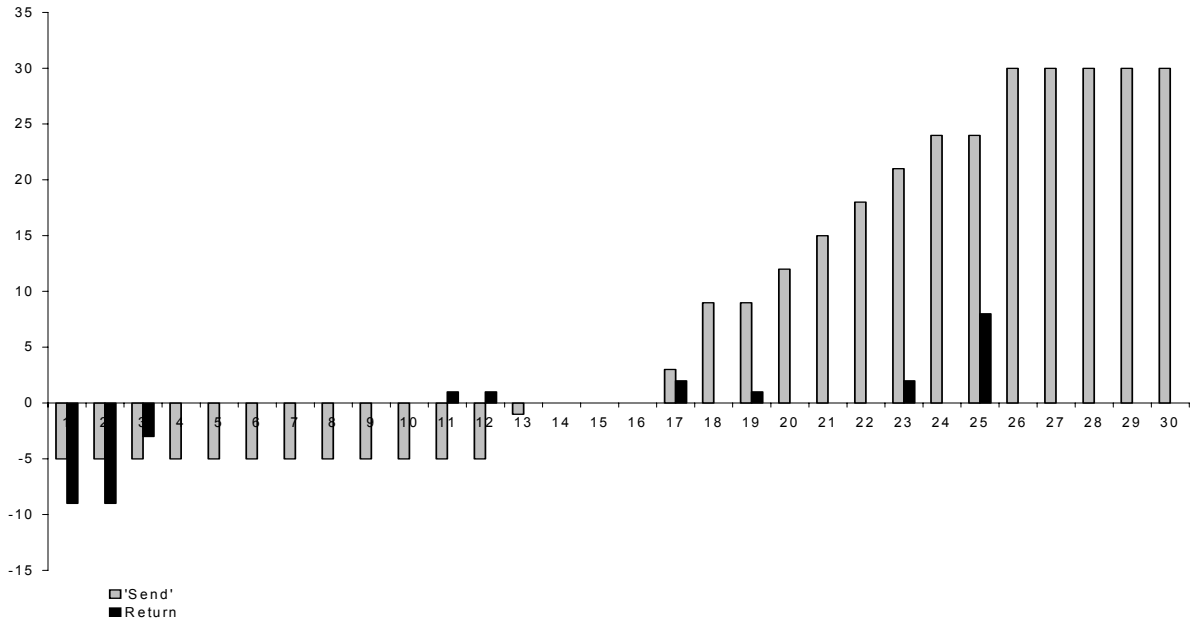


Figure 4. Money "Sent" and Returned in Treatment C

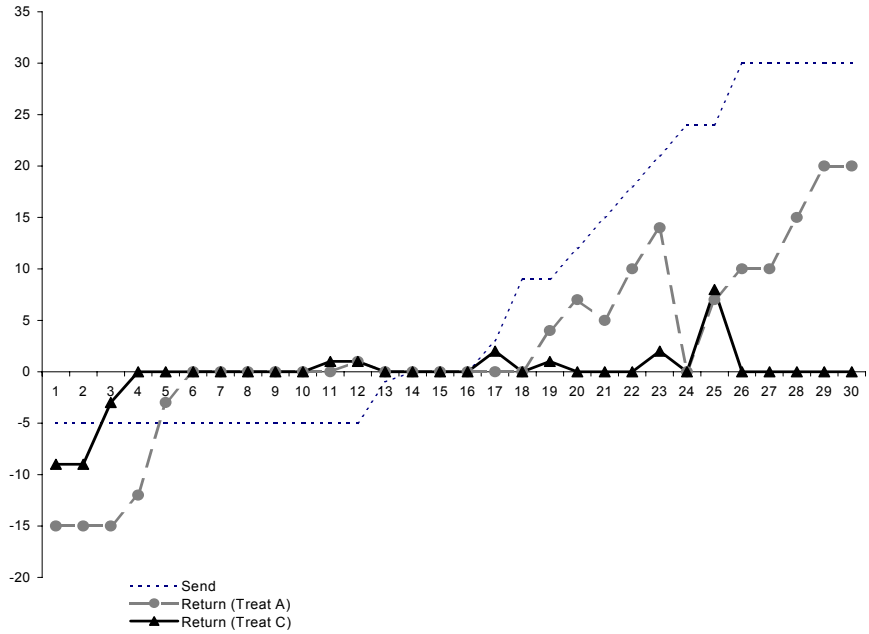


Figure 5. Money Sent in Treatment A and Returned in Treatments A and C