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 September 2001

Abstract

This paper uses a triadic experimental design to discriminate between actions motivated by 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 preferences. The triadic design includes the moonlighting game in which first-mover actions can motivate positively- or negatively-reciprocal actions by second movers. Thus, first movers can be motivated by trust (in positive reciprocity) or fear (of negative reciprocity) in addition to selfish, altruistic, or inequality-averse 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 in positive reciprocity and an absence of fear of negative reciprocity. Furthermore, the first movers' behavior is based on rational 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

Economic and game-theoretic modeling has 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 preference 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 actions of another. But actions that are inconsistent with self-regarding preferences can also be motivated by agents' altruistic or inequality-averse preferences over the distribution of material payoffs.[1] The distinction between actions motivated by 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 is to ask whether attribution of intentions is a significant motive for behavior in experimental 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 triadic experimental design to discriminate between actions motivated by 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). In the moonlighting game, a first-mover can either give money to a paired second mover or take money from the second mover. Any amount given is tripled by the experimenter. Any amount taken is not transformed by the experimenter. After a second mover learns of the tripled amount sent or the amount taken by the paired first mover, the second mover has an opportunity to give or take money from the first mover. Each dollar that the second mover gives to the first mover costs the second mover one dollar. Each dollar that the second mover takes from the first mover costs the second mover 33 cents.

Our starting point is to address the question of what motivations for actions can be inferred from observations in the moonlighting game. A first mover may give money to a second mover because of altruistic other-regarding preferences. Alternatively, a first mover may give money to the paired second mover because of trust that the second mover will return part of the profit from the experimenter's tripling of the amount sent. Furthermore, a first mover may refrain from taking money from the paired second mover because of altruistic or inequality-averse other-regarding preferences. Alternatively, a first mover may refrain from taking money from the

paired second mover because of fear that the second mover will retaliate by subsequently taking money from him. Turning now to the second mover, we note that a second mover may return part of the tripled amount sent by the paired first mover because of positive reciprocity, by which we mean a motivation to adopt a costly action to benefit someone whose intentional behavior has benefited oneself. Alternatively, the second mover may return part of the tripled amount sent because of altruistic or inequality-averse other-regarding preferences. Next consider possible second-mover motivations for incurring the cost of taking money from a first mover that took money from the second mover. The second mover may take money from the first mover because of negative reciprocity, by which we mean a motivation to adopt a costly action to inflict harm on someone whose intentional behavior has inflicted harm on oneself. Alternatively, the second mover may take money from the first mover because of inequality-averse other-regarding preferences.

We have noted that observations in the moonlighting game do not discriminate between trust and altruistic other-regarding preferences of first movers, nor between fear and inequality-averse other-regarding preferences of first movers, nor between positive reciprocity and altruistic or inequality-averse other-regarding preferences of second movers, nor between negative reciprocity and inequality-averse other-regarding preferences of second movers. This is our reason for incorporating the moonlighting game into a triadic experimental design that includes dictator control treatments that make these discriminations possible. The triadic design is described in section 3.

2. Definitions

Interpretations of data in this paper will be based on the following definitions. Preferences over the absolute and relative amount of another individual's money payoff, in addition to one's own money payoff, will be referred to as other-regarding preferences. Such preferences can involve ideas of the fairness of outcomes. For ease of explanation, consider the special case of two

agents. Let y^j and y^k denote the money payoffs of agents *j* and *k*. Assume that agent *k*'s preferences can be represented by a utility function. Then agent *k* has other-regarding preferences for the income of agent *j* if his/her utility function, $u^k(y^k, y^j)$ is *not* a constant function of y^j. Such preferences can be altruistic or inequality-averse. The preferences are altruistic if the utility function, u^K is globally increasing in both arguments, although perhaps having the property of egocentricity (Cox, Sadiraj, and Sadiraj, 2001). The preferences are inequality-averse (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) if u^K is increasing in y^k, decreasing in $\max(0, y^j - y^k)$, and possibly also decreasing in $\max(0, y^k - y^j)$. It is important to distinguish actions motivated by other-regarding preferences from actions motivated by trust, fear, or reciprocity.

The concepts of trust and positive reciprocity used in this paper are defined as follows. Agent 1 undertakes an action that exhibits trust if the chosen action: (a) creates a *monetary* gain that could be shared with agent 2; and (b) exposes agent 1 to the risk of a loss of *utility* if agent 2 defects and appropriates too much of the monetary gain. Agent 2 undertakes an action that exhibits positive direct reciprocity if the chosen action: (a) follows a trusting action by agent 1; (b) gives agent 1 a monetary gain; and (c) is undertaken instead of an available alternative action that would produce outcomes preferred by agent 2 in the absence of the trusting action by agent 1.

The concepts of fear and negative reciprocity used in this paper are defined as follows. Agent 1 undertakes an action that exhibits fear of negative reciprocity if, in two otherwise-identical environments, he: (a) foregoes an opportunity to take money from agent 2 when agent 2 has an opportunity to retaliate; and (b) takes money from agent 2 when agent 2 does not have an opportunity to retaliate. Agent 2 undertakes an action that exhibits negative direct reciprocity if in two otherwise-identical environments in which agent 1 has been allocated the higher monetary payoff he: (a) reduces *both* agents' monetary payoffs when agent 1 is responsible for the

unfavorable allocation; and (b) does not reduce their monetary payoffs when agent 1 is not responsible for the unfavorable allocation.

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 3 makes it possible to discriminate between the implications of other-regarding preferences and trust, fear, or reciprocity.

3. Experimental Design and Procedures

The experimental design includes the moonlighting game and two specially-designed dictator games.

3.1 The Three Games

The moonlighting game is experimental treatment A. In treatment A, a first mover chooses an amount, s^a to send to the second mover, where $s^a \in S$ and

(1) $S = \{-5,-4,\cdots,0,1,\cdots,10\}$.

If $s^a < 0$ then the first mover takes money from the second mover. If $s^a > 0$ then the first mover gives money to the second mover. Thus, if $s^a < 0$ then the second mover's money payoff is decreased by the amount $-s^a$ and the first mover's money payoff is increased by the amount $-s^a$. If $s^a > 0$ then the second mover's money payoff is increased by $3s^a$ and the first mover's money payoff is decreased by s^a. If $s^a = 0$ then no money payoff is changed.

The choice of s^a by the first mover selects the $\Gamma(s^a)$ subgame, in which the second mover chooses $r^a \in R(s^a)$, where

(2) $R(s^a) = \{m(s^a), m(s^a)+1, \cdots, M(s^a)\}$.

If $r^a < 0$ then the first mover's money payoff is decreased by $-3r^a$ and the second mover's money payoff is decreased by $-r^a$. If $r^a > 0$ then the first mover's money payoff is increased by r^a and the second mover's money payoff is decreased by r^a.

The second mover is not allowed to choose a return amount that would leave the first mover with a negative monetary payoff; therefore, $r^a \geq -((10 - s^a)/3))$. Also, r^a must be an integer; therefore,

(3) $m(s^a) = -\max(y \in Z \ni y \leq ((10 - s^a)/3))$,

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 $r^a \leq M(s^a)$, where

(4) $M(s^a) = 10 + s^a$, if $s^a < 0$

$= 10 + 3s^a$, otherwise.

The set, *S* of feasible choices for a first mover was chosen because it is a natural extension of the set of feasible choices in the investment game. The set of integers weakly between 0 and 10 is the feasible set of first-mover choices in the Berg, Dickhaut, and McCabe (1995) investment game experiment. Our choice of *S* for the moonlighting game extends the set of feasible choices to include the possibility of taking any integer amount up to 5, which is one-half of the second mover's endowment. Thus we can address the question of the effect on first-mover decisions of adding the opportunity to take money by comparing our data to the Berg, et al. data.

Limiting the amount that can be taken by a first mover to be no more than 5 preserves the ability of a second mover to retaliate if he chooses to. For example, if a first mover takes the

maximum of 5, then the second mover has the option of paying a cost in integer amounts up to all of his remaining balance of 5. If a first mover takes 5 from the second mover, then he has 15 before the second mover's decision. The second mover then has the option of paying all of his remaining 5 to reduce the first mover's credit balance to zero. The feasible choice sets for the other treatments are as follows.

Treatment B is a dictator game with the same strategy set for the "first mover" as in treatment A, the moonlighting game. Thus, the "first mover" chooses $s^b \in S$, where S is defined in statement (1). The "second mover" has no decision to make.

Treatment C is a dictator game in which the "first mover" has no decision and the "second mover" has the same strategy set that he has in the subgame selected in treatment A. Thus, the "second mover" chooses $r^c \in R(s^a)$, where $R(s^a)$ is defined in statement (2).

3.2 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 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 experimenter 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 one of the two parts of the laboratory were completely 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; the printed copies of the instructions were made available in case the subjects 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 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 and then asked to try again. After all of the subjects answered all questions correctly, the decision-making part of the experiment began.

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.

In treatment A, each individual in each group was credited with 10 euros. At the time of the experiment, one euro was worth a little less than one dollar. Each individual in group X was given the opportunity to give or take money from an anonymously-paired person in group Y. Thus, a group X subject could choose neither to give nor take. He could choose to give any integer amount up to all 10 of his euros to the group Y subject. Or he could take any integer amount up to 5 euros from the group Y subject. Any amount given was tripled by the experimenters. Any amount taken was not transformed by the experimenters. Subsequently, each individual in group Y was given the opportunity to give or take money in integer amounts from the anonymously-paired person in group X. Each euro given to the group X subject cost the group Y subject one euro. The amount given could not exceed the amount that would leave the group Y subject with zero euros. Each three euros taken from the group X subject cost the group Y subject one euro. The amount taken could not exceed the amount that would leave the group X subject with zero euros.

Treatment B differed from treatment A only in that the individuals in group Y had no decision to make. Thus, treatment B was a dictator experiment in which group X subjects had the same strategy set as in treatment A.

Treatment C differed from treatment A as follows. Individuals in group X had no decision to make. Each subject pair, $j = 1,2,\ldots,30$, 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 $10 + B_j$. 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.[3] 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 euros 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. The instructions given to the subjects are contained in an appendix available upon request to the authors. 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.

4. Subjects' Behavior in the Experiment

The subjects' behavior in treatment A, the moonlighting game is presented in Figure 1. 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 euros 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 euros. On average, it was profitable 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 euros 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 euros 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 the Moonlighting Game vs. Behavior in the Investment Game

The first-mover data from the (moonlighting game) treatment A can be compared with first-mover data from the "no history" investment game treatment in Berg, Dickhaut, and McCabe (1995). In their no history treatment, 30 out of 32 subjects gave positive amounts of money to the second movers and the mean amount given was $15.48 (or the mean amount sent, before multiplication by three, was $5.16). These data appear quite different from the first-mover data from our treatment A. Table 1 presents contingency tables comparing first- and second-mover data from their experiment with data from our moonlighting game treatment. Chi-square contingency table tests are used because the two experiments involve different strategy sets.

The contingency table in Table 1 for amounts sent by first-movers includes two classes: class 1 is the number of observations for which the amount sent was nonpositive and class 2 is the number of observations for which the amount sent was positive. The test reveals a significant

difference ($p < .001$) between first-mover data from the two experiments. Thus the extension of the investment game's feasible choice set to include opportunities for agents to take as well as give money has a very significant effect on first movers' decisions.

Next consider amounts returned. Let *S* denote the amount sent (measured by the cost to the first mover) and *R* denote the amount returned (measured by the cost to the second mover). The contingency table in Table 1 for amounts returned by second movers includes two classes: class 1 is the number of observations for which $S > 0$ and $R \geq S$; and class 2 is the number of observations for which $S > 0$ and $R < S$. Class 1 includes subject pairs in which the first mover gave money and the second mover shared the profit on the first mover's "investment," leaving both subjects with a profit. Class 2 includes subject pairs in which the first mover gave money and the second mover defected, leaving the first mover with a loss. The test does not reveal a significant difference between return behavior in the two experiments ($p > .1$). Thus the extension of the feasible choice set, to include the opportunity for both first and second movers to take money, does not significantly change the behavior of second movers who receive positive transfers.

4.2. 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 for the first movers to either give or take money. But the average loss from taking was larger than the average loss 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.).[4]

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 2 presents contingency tables comparing first- and second-mover data from their 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 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.

As above, let S denote the amount sent and R denote the amount returned (both measured by the cost to the decision-maker). The contingency table for amounts returned by second-movers includes four classes: class 1 is the number of observations for which $S \leq 0$ and $3R \leq S$; class 2

is the number of observations for which $S \leq 0$ and $\min(3R, R) > S$; class 3 is the number of observations for which $S > 0$ and $R \geq S$; and class 4 is the number of observations for which $S > 0$ and $R < S$. Note that class 1 includes subject pairs in which the first mover took money and the second mover strongly retaliated, leaving both subjects with a loss. Class 2 includes subject pairs in which the first mover took money and the second mover retaliated weakly or not at all, leaving the first mover with a profit. Class 3 includes subject pairs in which the first mover gave money and the second mover shared the profit on the first mover's "investment," leaving both subjects with a profit. Class 4 includes subject pairs in which the first mover gave money and the second mover defected, leaving the first mover with a loss.

The test reported in Table 2 reveals a significant difference between second-mover data from the two experiments ($p < .025$). Inspection of the contingency table reveals notable differences in the number of observations in classes 2 and 4. The larger number of observations in class 2 for our data is consistent with there being less negative reciprocity in our experiment than in theirs. The larger number of observations in class 4 for their data is consistent with there being less positive reciprocity in their experiment than in ours.[5]

The source of the difference in second-mover behavior is an interesting question. There were notable differences between the protocols used in the two experiments. As explained above, our experiment was computerized and involved double-blind payoff and group-separation procedures that created high "social distance" between the first- and second-mover groups of subjects and between the subjects and the experimenters. In contrast, the Abbink, et al. experiment was a "mensa" (cafeteria) experiment in which subjects participated in a manual (non-computerized) experiment as they entered the lobbies of two buildings. Abbink, et al. state that their payoff procedure was "double blind" *because* the subjects were instructed to use pseudonyms. But the description of their procedures does not specify whether the subjects were paid face-to-face by the experimenters, nor does it contain other details that would reveal the level of social distance

between the subjects and the experimenters in their protocol.[6] Of course, this discussion of experimental protocols should not be misinterpreted as a view that there is only one proper way to conduct a fairness experiment. To the contrary, running experiments with protocols involving different levels of social distance can yield deeper insight into fairness behavior, most especially about the extent to which social norms for reciprocity and other fairness-oriented behavior are internalized.

4.3 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 to neither take nor give money (i.e., return zero) 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 euros (i.e. send minus five) from the paired second mover.

As noted above, 12 out of the 30 first movers took five euros 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. Confining our attention to pairs of subjects, we observe that six of the first movers who took the maximum of five euros 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 3 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 3 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 and treatment C is presented in Figure 2 and Figure 3. 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 euros, four subjects took smaller amounts, and two subjects gave positive amounts. In treatment C, 21 out of 30 subjects chose zero euros, six subjects "returned" positive amounts, and three subjects "returned" negative amounts. The second and third rows of Table 3 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 3 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.

4.4 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 2 presents the amounts sent in treatments A and B. The third row of Table 4 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 in the positive reciprocity of second movers 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 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 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 3 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.

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 of negative reciprocity or trust in positive reciprocity. A first mover might prefer to take money from the paired second mover. If so, he will take money in treatment B; but he may not also take money in treatment A if he is

afraid of retaliation (i.e., negative reciprocity). 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. As seen in Figure 2, 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 2 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 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 profit from the tripling of amounts sent. As seen in Figure 2, 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 positive reciprocity 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.

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 1

and 3 show how second movers responded to amounts they received in treatments A and C. Figure 4 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 provided 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 euros. 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 euros to a high of 20 euros. Finally, note that the fourth row of Table 4 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 euros taken from them and the other subject had one euro taken. How did they 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 4 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.

5. 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 in the positive reciprocity of second movers. Furthermore, this trust was based on rational 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 euros 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 rational 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 euros after the second movers' decisions. But this conclusion about the rational 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. 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, 2001). 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. A model that includes both intentions and outcome preferences has been developed by Falk and Fischbacher (1999).

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, 2001), and the punishment 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 (2001) find that negative reciprocity is not significant in the punishment game and that positive reciprocity is significant in the trust game with a single-blind protocol but insignificant with a double-blind protocol.

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. We thank Thorsten Giertz and Jens Grosser for providing help in running the experiment.

1. Models of inequality-averse preferences are developed and applied in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000).
2. Models of (unconditional) preferences over outcomes are presented in Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Andreoni and Miller (2000), and Cox, Sadiraj, and Sadiraj (2001). 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 (1998) and Charness and Rabin (2000).
3. See Dufwenberg, Gneezy, Güth, and van Damme (2001) for tests of both direct and indirect reciprocity in the context of the investment game.
4. 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.
5. Excluding the observation excluded by Abbink, et al. leads to the following: (a) the entry in the second row and fourth column of Table 2 changes from 1 to 0; (b) the chi-square test statistic changes from 10.14 to 12.78; and (c) the p-value changes from $p < .025$ to $p < .01$.

6. 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 (2001) 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," Discussion paper, University of Wisconsin, 2000.

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." Discussion paper, University of California at Berkeley, 2000.

Cox, James C., "On the Economics of Reciprocity." Discussion paper, University of Arizona, January 2001.

Cox, James C. and Cary A. Deck, "On the Nature of Reciprocal Motives." Discussion paper, University of Arizona, September 2000; revised August 2001.

Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj, "A Theory of Competition and Fairness without Inequality Aversion." Discussion paper, University of Arizona and University of Amsterdam, January 2001; revised August 2001.

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.

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

	Send Data Contingency Table		Return Data Contingency Table	
	Class 1 ($S \leq 0$)	Class 2 ($S > 0$)	Class 1 ($S > 0{,}3R \geq S$)	Class 2 ($S > 0{,}3R < S$)
Treatment A	$16_{8.71}$	$14_{21.29}$	$10_{7.64}$	$4_{6.36}$
Berg, et al.	$2_{9.29}$	$30_{22.71}$	$14_{16.36}$	$16_{13.64}$
Chi-Square Test	16.66 ($p<.001$)		2.36 ($p>.1$)	

Table 2. Treatment A Data vs. Abbink, et al. Data

	Send Data Contingency Table		Return Data Contingency Table			
	Class 1 ($S \le 0$)	Class 2 ($S > 0$)	Class 1 ($S \le 0, 3R \le S$)	Class 2 ($S \le 0, \min(3R, R) > S$)	Class 3 ($S > 0, R \ge S$)	Class 4 ($S > 0, R < S$)
Treatment A	$16_{13.07}$	$14_{16.94}$	$7_{8.23}$	$9_{4.84}$	$10_{9.68}$	$4_{7.26}$
Abbink, et al.	$11_{13.94}$	$21_{18.07}$	$10_{8.77}$	$1_{5.16}$	$10_{10.32}$	$11_{7.74}$
Chi-Square Test	2.26 (p>.1)		10.14 (p<.025)			

Table 3. 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 4. 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		… …		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.

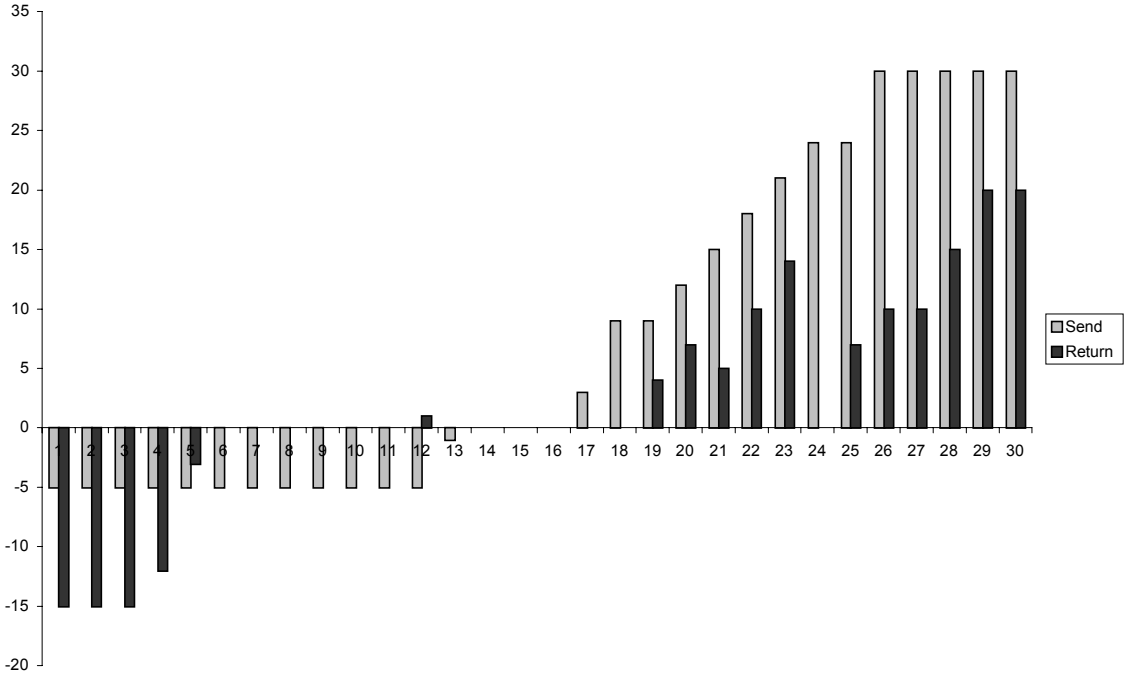

Figure 1. Money Sent and Returned in Treatment A

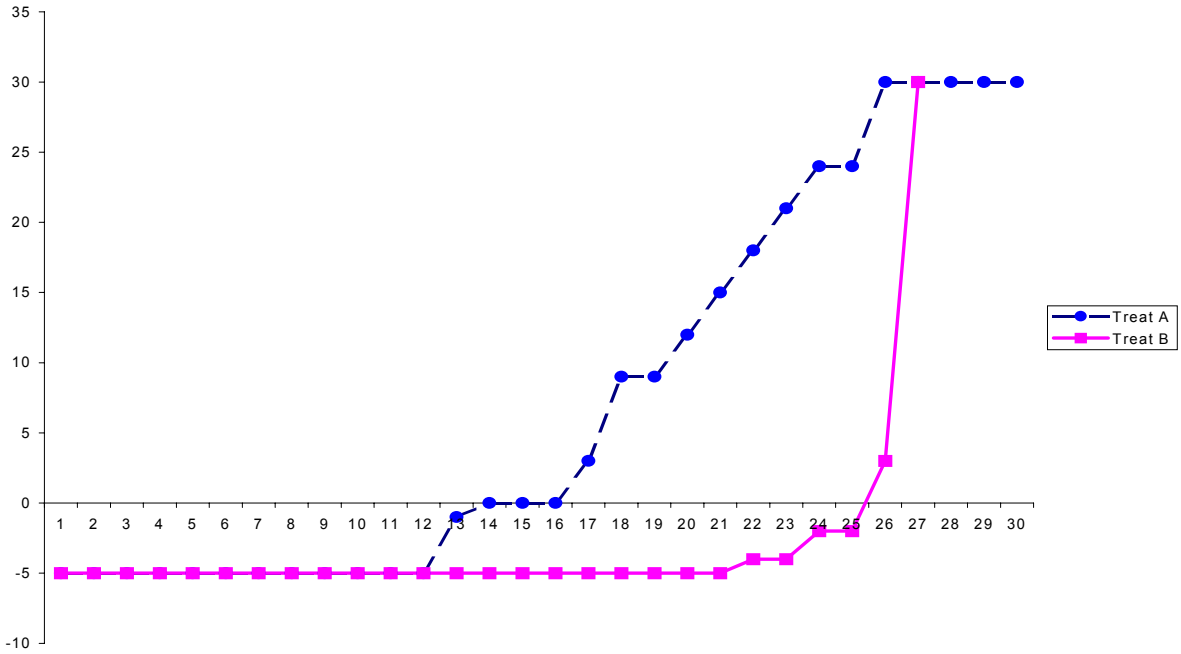

Figure 2. Money Sent in Treatments A and B

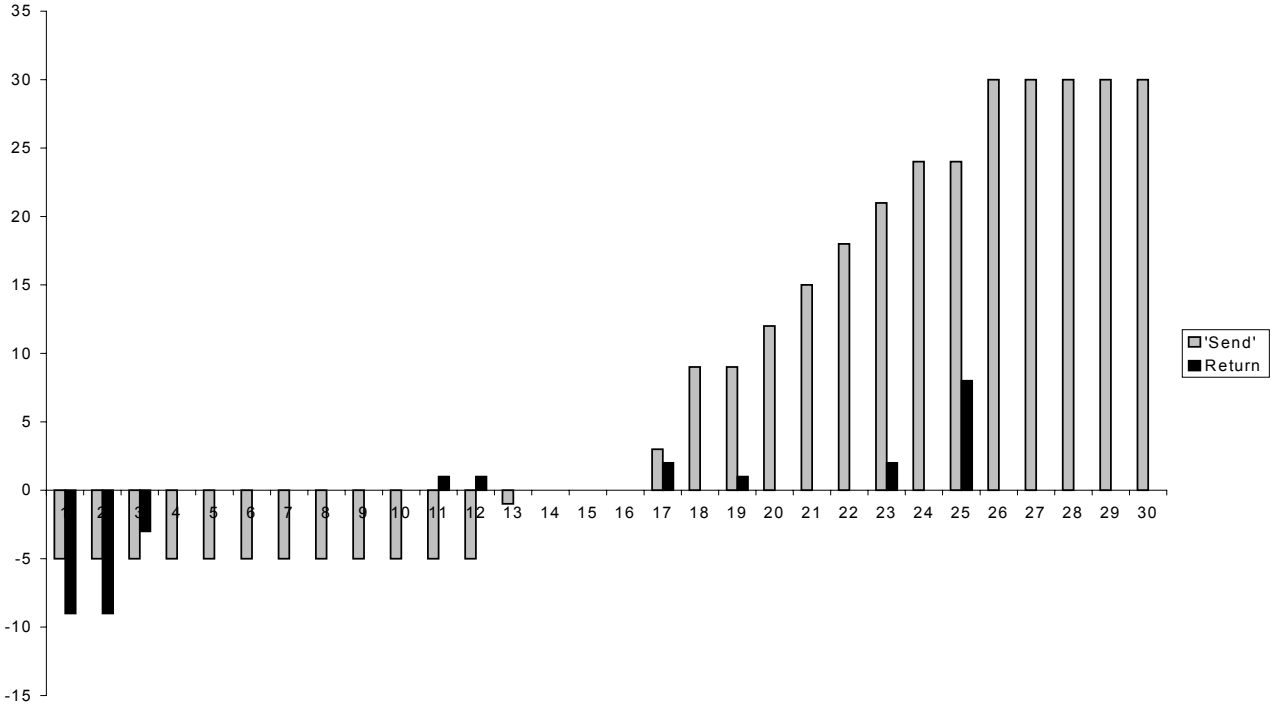

Figure 3. Money "Sent" and Returned in Treatment C

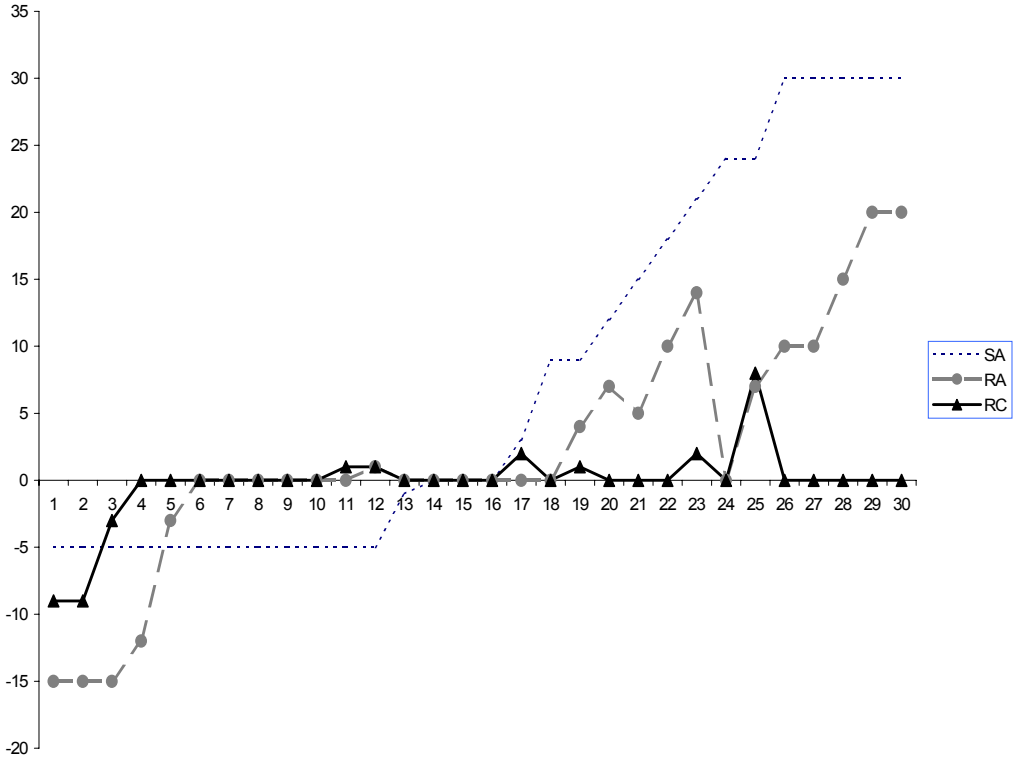

Figure 4. Money Sent in Treatment A and Returned in Treatments A and C