

Implications of trust, fear, and reciprocity for modeling economic behavior

James C. Cox · Klarita Sadiraj · Vjollca Sadiraj

Revised: 7 November 2006 / Accepted: 20 November 2006 /
Published online: 23 March 2007
© Economic Science Association 2006

Abstract This paper reports three experiments with triadic or dyadic designs. The experiments include the moonlighting game in which first-mover actions can elicit positively or negatively reciprocal reactions from second movers. First movers can be motivated by trust in positive reciprocity or fear of negative reciprocity, in addition to unconditional other-regarding preferences. Second movers can be motivated by unconditional other-regarding preferences as well as positive or negative reciprocity. The experimental designs include control treatments that discriminate among actions with alternative motivations. Data from our three experiments and a fourth one are used to explore methodological questions, including the effects on behavioral hypothesis tests of within-subjects vs. across-subjects designs, single-blind vs. double-blind payoffs, random vs. dictator first-mover control treatments, and strategy responses vs. sequential play.

Keywords Experiments · Theory · Parsimony · Trust · Fear · Reciprocity · Methodology

Electronic Supplementary Material Supplementary material is available in the online version of this article at <http://dx.doi.org/10.1007/s10683-006-9156-7>.

J. C. Cox (✉)

Department of Economics and Experimental Economics Center (ExCen), Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA 30302-3992
e-mail: jccox@gsu.edu

K. Sadiraj

Social and Cultural Planning Office of The Netherlands (SCP), PO Box 16164, 2500 BD, The Hague, The Netherlands
e-mail: K.Sadiraj@scp.nl

V. Sadiraj

Department of Economics and Experimental Economics Center (ExCen), Andrew Young School of Policy Studies, Georgia State University, Atlanta, GA 30302-3992
e-mail: vsadiraj@gsu.edu

JEL Classification C70, C91, D63, D64

1 Introduction

Economics traditionally focused on the model of self-regarding (or “economic man”) preferences in which agents were assumed to be exclusively concerned with their own material payoffs. Actions that are inconsistent with the predictions of this model can be motivated by a desire to reciprocate the actions of another. Kind actions by one person can elicit positively reciprocal reactions from another. Negatively reciprocal reactions can be elicited by unkind actions. Anticipation of another’s response, such as trust by one person in the positive reciprocity of another or fear by one person of the negative reciprocity of another, can also lead to behavior inconsistent with the self-regarding preferences model. Alternatively, actions that are inconsistent with self-regarding preferences can be motivated by an agent’s (unconditional) altruistic or inequality-averse other-regarding preferences rather than trust, fear, or reciprocity.

Recent theoretical developments reflect the distinction between models of unconditional preferences and models incorporating reciprocity. Unconditional preferences models include the Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) inequality aversion models, the Charness and Rabin (2002) quasi-maximin (distributional) model, and the Andreoni and Miller (2002) and Cox and Sadiraj (2004) altruism models. A model that incorporates reciprocity and status into preferences is developed by Cox et al. (in press) and a nonparametric generalization of this reciprocal preferences approach is reported by Cox et al. (2006).

Unconditional preferences models are relatively easy to apply to data and should, by the principle of parsimony in theoretical modeling, be preferred to more complicated, conditional preferences models in the absence of convincing evidence of the empirical significance of reciprocity as a determinant of behavior. The empirical question that is central to assessing the implications of the parsimony principle of scientific theorizing is whether anticipations or reactions can be shown to be significant determinants of behavior. If many people reveal other-regarding preferences, then are these revealed preferences independent of the anticipated or observed actions of others or conditional on them? In what environments and game forms do people reveal preferences that are independent of past actions or anticipated future actions of others? What environments and game forms, if any, can be demonstrated to elicit responses from people that depend in identifiable ways on the observed or anticipated actions of others?

In order to obtain data that can guide development of economic models that are consistent with behavior, and apply the principle of parsimony for choosing among models, we need to discriminate among actions with alternative motivations. We use a three-games or triadic experimental design including the moonlighting game and two dictator games to discriminate between behavior that can be modeled with unconditional preferences over outcomes and behavior that requires introduction of anticipation of others’ future actions and/or reactions to their past actions.

As in the Abbink et al. (2000) experiment with the moonlighting game and in the Berg et al. (1995) experiment with the investment game, our first experiment uses a double blind payoff protocol in which neither other subjects nor the experimenter knows the identity of a subject who makes any specific decision. Our first experiment also compares behavior in the moonlighting game with two dictator control treatments

“across subjects”; that is, different subjects are used in the three treatments. We use alternative protocols to explore several methodological questions, including the effects of (a) within-subjects vs. across-subjects designs, (b) single-blind vs. double-blind payoffs, and (c) random first-mover control and strategy method vs. dictator first-mover control and sequential decisions. Data from our three experiments are used to explore questions (a) and (b). Data from our experiments together with data reported by Falk et al. (2000) are used to address question (c). The Falk et al. paper reports a dyadic design experiment with the moonlighting game and random selection of first moves as a control treatment for second mover behavior.¹

There are a few other studies that use control treatments for motivations. Blount (1995), Charness (2004), and Offerman (2002) use random selection of first moves as a reactions-control treatment for second mover behavior in various games; Falk et al. (2000) use this design for the moonlighting game. Bohnet and Hong (2004) use random selection of second moves as an anticipations-control treatment for first mover behavior in a trust game. Bolton et al. (1998) use a reactions-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 uses dictator control treatments for both second movers’ reactions to first movers’ actions and first movers’ anticipations of second movers’ reactions. Cox (2002, 2004) uses control treatments for both reactions and anticipations in experiments with the investment game in which there can be positive reciprocity and trust in positive reciprocity. The present paper extends 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 testable hypotheses

The experimental design uses three treatments. Treatment A is the central game of interest, the moonlighting game. Treatments B and C are specially designed dictator games that provide control treatments for first mover and second mover motivations in the moonlighting game.

2.1 The three treatments

In Treatment A, the first mover chooses a feasible set for the second mover by choosing an amount to give to or take from the second mover. The second mover is a dictator who chooses an allocation in this feasible set that determines both first mover and second mover money payoffs. Each second mover is credited with a money endowment of 10 (dollars or euros). Each first mover is credited with a money endowment of 10 and given the task of deciding whether she wants to give to a paired second mover none, some, or all of her endowment or take up to 5 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

¹ The experiment in Falk et al. (2000) and our Experiment 1 were run independently of each other, theirs in 1998 and ours in 2000.

money from her. Each dollar (or euro) that the second mover gives to the paired first mover costs the second mover 1 dollar (or euro). Each three dollars (or euros) that the second mover takes from the paired first mover costs the second mover one dollar (or 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.

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. In Treatment B the first mover has the same feasible set of choices as in Treatment A. In Treatment B, however, the first mover's choice determines the money payoffs of both subjects rather than determining the second mover's feasible set.

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. The dictator, "second mover" is given one of the feasible sets determined by a first mover's choice in Treatment A (but the dictator does not know what determined the feasible set). The dictator then determines both subjects' money payoffs by choosing an integer amount to give to or take from the paired subject.

2.2 Tests for anticipations and reactions

The triadic design provides data that can be used to discriminate empirically among choices determined solely by unconditional distributional preferences and choices determined in part by reaction to another's previous action or anticipation of her possible future action. The specific ways in which data generated with the triadic design can be used in empirical applications of theoretical models of social preferences that incorporate reciprocity are explained in Cox et al. (in press) and Cox et al. (2006). A detailed explanation of how data from triadic-design experiments with the moonlighting game support tests for anticipations, such as trust and fear, and reactions such as positive and negative reciprocity is contained in appendix 1 on the journal's web site; we here present a summary explanation.

Treatment B differs from Treatment A only in that the "second mover" does not have a decision to make; thus he does not have an opportunity either to take money from the "first mover" or give money to her. Since a "second mover" has no action that can be taken in Treatment B, a first mover's choice cannot be affected by anticipations of what that action might be. In Treatment A, in contrast, a second mover can give or take money from the first mover after observing the amount of money the first mover has given to or taken from the second mover. Hence in making her decision in Treatment A, a first mover's choice can be affected by anticipations of how the second mover will react: the first mover may trust that the second mover will positively reciprocate a positive transfer or fear that the second mover will negatively reciprocate a negative transfer. Of course, a first mover may also anticipate that a second mover will return or take money because of non-reciprocal altruistic or inequality-averse preferences. Conclusions about whether first mover actions in the moonlighting game (Treatment A) are motivated by anticipations such as trust or fear can be supported by observations of the difference between Treatments A and B in the amounts of money first movers give to or take from second movers. If the amount s^a sent by the first mover in Treatment A is larger than the amount s^b sent in Treatment B then anticipation of second mover

reaction has been revealed to have an effect on first mover behavior. A testable hypothesis about an anticipations effect is stated in the top, left entry of Table 1. If the amount s^a sent in Treatment A is positive, and greater than the (positive, zero, or negative) amount s^b sent in Treatment B, then the choices in Treatments A and B together support the conclusion that the first mover's anticipation is trusting. A testable hypothesis about trust is stated in the top, middle entry in Table 1. If the amount sent in Treatment B is negative, and the nonpositive amount transferred in Treatment A is larger, then choices in Treatments A and B together support the conclusion that the first mover's anticipation is fearful. A testable hypothesis about fear is stated in the top, right entry in Table 1.

Treatment C differs from Treatment A only in that the "first mover" does not have a decision to make; thus he does not have an opportunity either to take money from the "second mover" or to give money to her. Since a "first mover" has no action that can be taken in Treatment C, a second mover's choice cannot be affected by reactions to a first mover's choice. In Treatment A, in contrast, a second mover observes an actual choice by a first mover to give to or take money from the second mover. Hence a second mover's choice in Treatment A can be affected by reaction to the choice by the first mover: the second mover may positively reciprocate a positive transfer or negatively reciprocate a negative transfer. Of course, a second mover may also give to or take money from a first mover because of non-reciprocal altruistic or inequality-averse preferences. Conclusions about whether second mover actions in the moonlighting game (Treatment A) are motivated by reactions such as positive reciprocity or negative reciprocity can be supported by observations of the difference between Treatments A and C in the amounts of money second movers give to or take from first movers. If the absolute value of the amount r^a returned by the second mover in Treatment A is larger than the absolute value of the amount r^c returned in Treatment C then reaction to the first mover's choice has been revealed to have an effect on second mover behavior. A testable hypothesis about a reactions effect is stated in the bottom, left entry of Table 1. Given a positive transfer by a first mover in Treatment A, if the amount r^a returned in Treatment A is positive and greater than the amount r^c returned in Treatment B, then the choices in Treatments A and C together support the conclusion that the second mover's reaction is positively reciprocal. A testable hypothesis about positive reciprocity is stated in the bottom, middle entry in Table 1. Given a negative transfer by the first mover in Treatment A, if the amount returned in Treatment A is less than the amount returned in Treatment C, then choices in Treatments A and C together support the conclusion that the second mover's reaction is negatively reciprocal. A testable hypothesis about negative reciprocity is stated in the bottom, right entry in Table 1.

3 The three experiments

Each of the three experiments includes the moonlighting game, Treatment A and one or two dictator control treatments. Experiments differ from each other with respect to the payoff protocol used and whether the comparisons between treatments are across-subjects or within-subjects. Experiment 1 uses a double-blind payoff protocol and an across-subjects design, Experiment 2 uses a single-blind payoff protocol and a within-subjects design, and Experiment 3 uses a single-blind payoff protocol and across-subjects design. Three experimental designs are used to explore the robustness of

Table 1 Testable hypotheses for anticipations and reactions

| Anticipations | Trust | Fear |
|---------------------|------------------------------|-------------------------|
| | First mover behavior | |
| $H_o^A : s^a = s^b$ | $H_a^A : s^a > s^b$ | $(s^a \leq 0, s^b < 0)$ |
| | $H_o^T : s^a = \max(0, s^b)$ | $H_a^T : s^a = s^b$ |
| | $H_a^T : s^a > \max(0, s^b)$ | $H_o^F : s^a > s^b$ |
| Reactions | Positive reciprocity | Negative reciprocity |
| | Second mover behavior | |
| $H_o^R : r^a = r^c$ | $(s^a > 0)$ | $(s^a < 0)$ |
| | $H_a^R : r^a > r^c $ | $H_o^N : r^a = r^c$ |
| | $H_o^P : r^a = r^c$ | $H_a^N : r^a < r^c$ |
| | $H_a^P : r^a > r^c$ | |

Amount sent by a first mover is denoted by s .

Amount returned by a second mover is denoted by r .

Superscripts are a , b , and c for Treatments A, B, and C, respectively.

Table 2 Experimental designs and protocols

| | Payoff protocol | Method | Experimental design | Subject pool | No. of pairs (subjects) |
|----------------|-----------------|--------------------|---------------------|---------------|--|
| Experiment 1 | Double-blind | Sequential choices | Across-subjects | UvA students | 30 ^{ac} , 27 ^b (174) |
| Experiment 2 | Single-blind | Sequential choices | Within-subjects | UofA students | 33 ^{abc} (99) |
| Experiment 3 | Single-blind | Sequential choices | Across-subjects | UofA students | 32 ^a , 35 ^c (134) |
| Experiment FFF | Single-blind | Strategy choices | Across-subjects | UofZ students | 33 ^a , 23 ^c (112) |

Superscripts are a, b, and c for Treatments A, B, and C, respectively.

conclusions about trust, fear, and reciprocity. One question that we address is whether the conclusions differ for within-subjects and across-subjects designs. Another question is concerned with the effect, if any, of single-blind or double-blind payoff protocols on the conclusions. In Section 5, we further explore robustness questions by comparing our data to data from the across-subjects, strategy-method, single-blind experiment with the moonlighting game reported by Falk et al. (2000).

3.1 Experiment 1: Across-subjects, double-blind

Experiment 1 sessions were run with custom computer software in the CREED laboratory at the University of Amsterdam in the fall of 2000. Subjects were randomly selected from the data base of students registered with the CREED laboratory for participation in experiments. The experiment included three treatments implemented in an across-subjects design. The payoff protocol was double blind. All money payoffs and subjects' feasible choices were quoted in euros. At the time the experiment sessions were run, 1 euro was worth a little less than 1 dollar.² The central features of Experiment 1 are listed in the first row of Table 2.

The decision-making part of Experiment 1 proceeded as follows. Subjects were divided equally in two parts of the laboratory completely separated by a thick floor to ceiling partition. The experimenters did not enter the laboratory when subjects were present. The monitor computer randomly determined which part of the laboratory was the room with first mover subjects and which was the room with second mover subjects. First and second mover pairing of subjects was established by where the subjects sat in the two separated parts of the laboratory. 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 mailbox 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.

² At that time, the euro was not yet a circulating currency but prices in retail stores were quoted in both Guilders and euros. The subjects were paid in Guilders, using the official exchange rate of 2.20 Guilders per euro. The experiment used euros 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).

Each of Treatments A, B, and C was run in four distinct sessions. The treatments were implemented “across-subjects”; that is, different subjects participated in each of the three treatments. In total 87 different pairs of subjects participated in this experiment. 30 different subject pairs participated in each of Treatments A and C and 27 in Treatment B.

3.2 Experiment 2: Within-subjects, single-blind

Experiment 2 was run with pencil and paper in the Economic Science Laboratory at the University of Arizona in the fall of 2005. Subjects were randomly selected from the data base of students registered with ESL for participation in experiments after elimination of names of any students who had previously participated in similar “fairness experiments.” All money payoffs and subjects’ feasible sets were in dollars. The experiment included three treatments implemented in a within-subjects design and used a single-blind payoff protocol in which the identity of subjects making specific decisions was known to the experimenters but not to other subjects. Thirty-three groups, each with three subjects, participated in this experiment. The central features of Experiment 2 are listed in the second row of Table 2.

Subjects were randomly assigned to three person groups. Each group consisted of a type X person, a type Y person, and a type Z person. Subjects did not know which two of the other people in the lab were in the same group. The type X and Y subjects in a group made decisions that corresponded to the Treatments A, B, and C decisions in Experiment 1. The type Z subjects had no decision to make. The subjects in each group participated in two rounds of the experiment. They were informed that one of the two rounds would be selected for money payoff by an experimenter flipping a coin in their presence.

3.3 Experiment 3: Across-subjects, single-blind

Experiment 3 included two treatments, the moonlighting game and a dictator game that provides a motivation control for second movers in the moonlighting game. Experiment 3 was run with custom computer software in the Economic Science Laboratory at the University of Arizona in the fall of 2005. As with Experiment 2, subjects were randomly selected from the data base of students registered with ESL after elimination of names of any students who had previously participated in similar experiments. The subject interface, feasible choices, and framing of subjects’ decision tasks were the same as Treatments A and C in Experiment 1, run at CREED. The subject instructions were an English translation of the Dutch instructions used at CREED in Experiment 1. All money payoffs and subjects’ feasible payoffs were quoted in numbers of dollars. This experiment used the same design of Treatments A and C as Experiment 1 but involved a single-blind payoff protocol. There were 32 pairs of subjects in Treatment A and 35 different pairs of subjects in Treatment C. The central features of Experiment 3 are listed in the third row of Table 2.

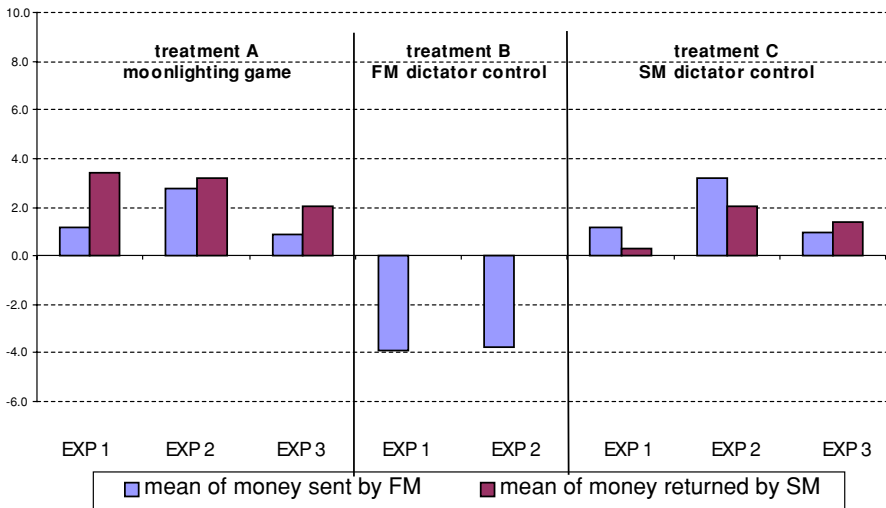


Fig. 1 Average amounts sent and returned by first movers and second movers

4 Subjects’ behavior in the three experiments

Figure 1 reports average amounts sent and returned by subjects in each of the three treatments in Experiments 1 and 2 and the two treatments in Experiment 3. Treatment labels are at the top of the figure and experiment labels at the bottom. The scale of numbers of euros or dollars is on the vertical axis of Fig. 1. Lightly-shaded bars report first mover (FM) data and darkly-shaded bars report second mover (SM) data. For example, the left-most pair of bars show that the average amount sent by FMs in Experiment 1 was 1.13 euros and the average amount returned by SMs in Experiment 1 was 3.43 euros. The three pairs of bars in the Treatment A (left) panel of Fig. 1 show that, on average, SMs returned more than FMs sent in each of the three experiments. The two bars in the Treatment B (middle) panel of the figure show that, on average, dictators in the FM control treatment took about 4 euros in Experiment 1 and slightly less than 4 dollars in Experiment 2. The dark bars in the Treatment C (right) panel of Fig. 1 show the average amounts “returned” by dictators in the SM control treatments. Reflecting the properties of the experimental designs, the lightly-shaded bars showing amounts “sent” in Treatment C of Experiments 1 and 3 are the same as the lightly-shaded bars for Treatment A (because in both treatments these are the amounts chosen by FMs in Treatment A). In contrast, the lightly-shaded bar for Treatment C of Experiment 2 shows an average of amounts determined by the experimenters for this within-subjects-design experiment.

The rest of Section 4 is concerned with using data from pairs of treatments within experiments to test hypotheses in Table 1 about anticipations and reactions and to ask whether such reciprocal behavior is robust to the differences among the designs and protocols for the three experiments. Section 5 compares data for individual treatments across experiments in order to address several methodological questions.

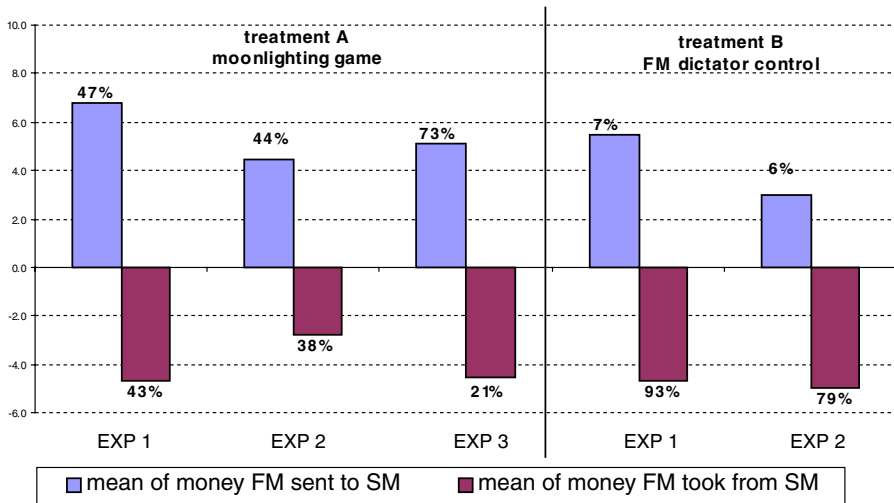


Fig. 2 Average amounts given and taken by first movers

4.1 Anticipated reactions are significant

Figure 2 shows a decomposition of FM choices in Treatments A and B into strictly positive and strictly negative amounts sent and the percentages of subjects making these choices. For example, the left-most pair of bars in Fig. 2 shows that, out of 30 FMs in Treatment A of Experiment 1: (a) 47% of them sent positive amounts that averaged 6.22 euros; (b) 43% of them sent negative amounts that averaged -4.69 euros; and (c) 10% ($= 100\% - 47\% - 43\%$) sent 0. In contrast, the bars for Treatment B show that only 7% of the subjects in Experiment 1 sent positive amounts, averaging 5.50 euros, and 93% of them sent negative amounts averaging -4.68 euros.³

The next question addressed is whether there are significant differences between FMs' choices in Treatment A (the moonlighting game) and Treatment B (the first-mover-control dictator game). The null Hypothesis H_o^A , that FMs send the same amounts in Treatments A and B, and alternative Hypothesis H_a^A , that they send more in Treatment A, are stated in the left column and top row of Table 1. As reported in the anticipations panel in Table 3, two nonparametric tests reject the null hypothesis in favor of the alternative hypothesis at 1% significance level for both experiments. Therefore, anticipated reactions of second movers are a significant determinant of first movers' decisions to give or take money. In the next two subsections we ask whether first movers' anticipations are trusting or fearful or both.

³ Behavior in Treatment B might, at first, seem to be unusual. The relation between models of social preferences and behavior in this dictator game and several other specially-designed dictator games is examined in detail in Cox and Sadiraj (2004).

Table 3 Tests for anticipations

| | Anticipations | | Trust | | Fear | |
|----------------------|-------------------------------|-----------|--|-----------|-------------------------------|---------|
| | (the amount of money FM sent) | | (the positive amount of money FM sent) | | (the amount of money FM took) | |
| | 1 | 2 | 1 | 2 | 1 | 2 |
| Tests for experiment | | | | | | |
| Nobs (Tr.A, Tr.B) | (30, 27) | (33, 33) | (30, 27) | (33, 33) | (16, 14) | (9, 9) |
| Tr:A Mean | 1.13 | 2.73 | 1.13 | 2.73 | -3.81 | -3.56 |
| (std. dev.) | (5.975) | (4.907) | (5.975) | (4.907) | (2.136) | (2.242) |
| Tr:B Mean | -3.93 | -3.73 | 0.41 | 0.18 | -5.00 | -5.00 |
| (std.dev.) | (3.100) | (2.516) | (1.927) | (0.769) | (0.000) | (0.000) |
| Smirnov | -0.526*** | -0.667*** | -0.396*** | -0.667*** | -0.250 | -0.333 |
| Epps-Singleton | 29*** | 50*** | 86*** | 105*** | 41.20**□ | -0.40□ |

***, ** and * denote significance at 1%, 5% and 10%, respectively.

We don't run *t*-tests here since data are not normally distributed.

□ Results may be inaccurate (because the matrix is close to singular or badly scaled).

Tr:A is the moonlighting game, Tr:B is the first-mover-control, dictator game.

4.2 Trusting behavior is significant

Experiment 2 was run with a within-subjects design, that is the same subjects made decisions in Treatments A and B. Comparing choices in the two treatments, we find that 73% of the subjects (24 out of the total of 33) made choices that reveal trust in Experiment 2. Since Experiment 1 was run across subjects, we report data at the aggregate level for both experiments. In Experiments 1 and 2, respectively, 47% (14 out of 30) and 73% (as reported above) of the subjects sent positive amounts of money to the second mover in the moonlighting game. In contrast, only 7% (2 out of 27) and 6% (2 out of 33) of the subjects in Experiments 1 and 2, respectively, sent positive amounts of money to the other person in the Treatment B dictator game. Data support a conclusion that subjects' behavior exhibits trust if amounts sent by FMs to paired subjects are positive and significantly larger in Treatment A than in Treatment B. Two nonparametric tests of the hypothesis in the top, middle panel of Table 1 reject the null Hypothesis H_o^A , that nonnegative amounts sent in Treatment A are equal to amounts sent in Treatment B, in favor of the alternative hypothesis of revealed trust H_o^T , that nonnegative amounts sent in Treatment A are larger, at 1% significance level for data from both Experiments 1 and 2, as reported in the middle panel of Table 3. We conclude that many FMs' anticipations in the moonlighting game are trusting in both Experiments 1 and 2; thus this central conclusion is robust to the alternative across-subjects and within-subjects experimental designs with, respectively, double-blind and single-blind payoff protocols.

4.3 Fearful behavior is of questionable significance

Next consider the question of whether FMs' behavior in Treatment A is characterized by fear of negative reciprocity. At the individual level, in Experiment 2, we observe that 12% (3 out of 26) of the subjects who took money in the dictator game made choices in the moonlighting game that reveal fear, that is they took money in both games but took less in the moonlighting game than in the dictator game. In Experiment 1, we observe that 16% (4 out of 25) of the subjects made choices that are consistent with fear. Table 3 presents results from tests using observations in which amounts sent in Treatment A are nonpositive and amounts sent in Treatment B are negative. The Smirnov test of the hypothesis in the top, right panel of Table 1 does *not* reject the null Hypothesis H_o^F , that amounts taken (the negative of amounts sent s , when $s < 0$) are the same in Treatments A and B, in favor of the alternative Hypothesis H_a^F , that amounts taken in Treatment A are smaller, for either Experiment 1 or Experiment 2, as reported in the right panel of Table 3. (The Epps-Singleton test is unreliable for this data, as described in a footnote to Table 3.) We conclude that FM's anticipations, that were *not* trusting, were also *not* fearful. This conclusion is supported by data from both across-subjects and within-subjects designs with, respectively, double-blind and single-blind payoffs.

4.4 Reactions are significant

Figure 3 shows a decomposition of SM data for Treatments A and C that are sorted by the criterion of whether the paired FMs sent positive or negative amounts in

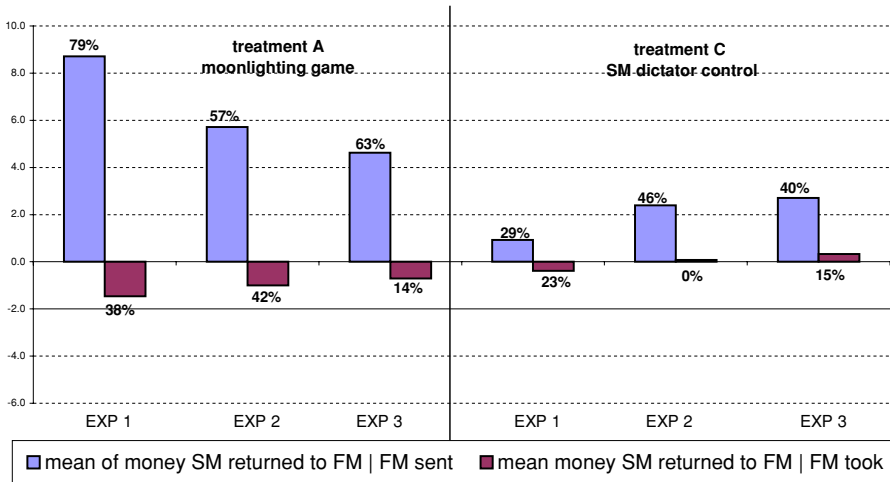


Fig. 3 Average second mover responses to positive and negative transfers

Treatment A. For example, the left-most pair of bars in Fig. 3 shows that in Treatment A of Experiment 1: (a) the average amount returned by SMs who received positive amounts was 8.71 euros and 79% of them returned strictly positive amounts; and (b) the average amount “returned” by SMs who “received” negative amounts was −1.46 euros and 38% “returned” strictly negative amounts. In contrast, Experiment 1 bars for Treatment C show that: (a) the average amount returned by SMs who had larger altered endowments than paired subjects was 0.93 euros and 29% of them gave strictly positive amounts; and (b) the average amount taken by SMs who had smaller endowments was −0.38 and 23% took strictly negative amounts.⁴

The question addressed next is whether there are significant differences between second movers’ choices in Treatment A (the moonlighting game) and Treatment C (the second-mover-control dictator game). Of course, an SM in the moonlighting game is also a dictator; the difference between the two games is that Treatment A dictators may treat the paired subjects differently than Treatment C dictators because of the intentional actions of the first movers in the former. We here test the null Hypothesis H_o^R , that amounts returned in Treatments A and C are equal, against the alternative Hypothesis H_a^R , that the absolute value of amounts returned is larger in Treatment A, as stated in the bottom, left panel of Table 1.

The left panel of Table 4 reports paired t -tests and paired Wilcoxon tests. The reported numbers of observations used in the paired tests are the numbers of distinct values for amounts sent by FMs in Treatment A (not the numbers of subjects whose responses are included in the test), except for within-subjects tests reported in the column marked with the # superscript. For example, the (11,11) entry for Nobs (Tr.A,Tr.C) for the reactions tests for Experiment 1 refers to 11 distinct values of

⁴ The dark bars for Treatment C of Experiments 2 and 3 each contain one possibly anomalous observation in which an SM who was allocated an endowment less than the paired subject chose to give that subject even more.

Table 4 Tests for reactions

| | Reactions (means of abs. amounts of money SM returned) | | | Positive reciprocity (means of amounts SM returned FM sent) | | | Negative reciprocity (means of amounts SM returned FM took) | | |
|-------------------------------------|---|----------------|----------------|--|----------------|----------------|--|----------------|----------------|
| | 1 [^] | 2 [#] | 3 [^] | 1 [^] | 2 [#] | 3 [^] | 1 ⁺ | 2 [#] | 3 [^] |
| Tests for experiment | | | | | | | | | |
| No. of ind. choices (Tr.A, Tr.C) | (30, 30) | (16, 16) | (33, 33) | (14, 14) | (12, 12) | (24, 24) | (12, 12) | (4, 4) | (6, 9) |
| Nobs for tests (Tr.A, Tr.C) | (11, 11) | (16, 16) | (7, 7) | (8, 8) | (12, 12) | (6, 6) | (12, 12) | (4, 4) | (6, 9) |
| Tr.A mean | 5.28 | 3.94 | 3.40 | 7.06 | 4.83 | 3.83 | -1.58 | -1.25 | -0.83 |
| (std.dev.) | (5.541) | (5.674) | (2.673) | (5.506) | (6.221) | (2.6523) | (2.392) | (2.500) | (2.041) |
| Tr.C mean | 0.81 | 1.88 | 2.21 | 1.06 | 2.25 | 2.53 | -0.42 | 0.75 | 0.33 |
| (std.dev.) | (1.311) | (2.277) | (1.787) | (1.474) | (2.417) | (1.734) | (1.311) | (1.500) | (1.000) |
| <i>t</i> -test (paired) | 2.576*** | 1.691* | 2.017** | 2.777*** | 1.683* | 1.906* | -1.482* | -1.633* | -1.486* |
| Wilcoxon (paired) | 2.010** | 1.409* | 1.690** | 1.960** | 1.506* | 1.572* | -1.230 | -1.400* | -1.394* |

***, ** and * denote significance at 1%, 5% and 10%, respectively.

[^]Tests in this column are paired according to value of amount sent *s*.

[#] Tests in this column are paired according to subject ID.

⁺Tests in this column are non-paired; they use amounts returned when amount "sent" is -5.

Tr.A is the moonlighting game and Tr.C is the second-mover-control, dictator game.

observed amounts sent or taken by FMs (in Treatment A) for which the mean amounts returned in Treatments A and C are paired. The null hypothesis is rejected by both tests at 5% significance with Experiment 1 and 2 data; for Experiment 3 data, the t -test is significant at 6% (and hence at 10%) and the Wilcoxon test is not significant.

In addition, for Experiment 2 data one can run within-subjects tests. The 2[#] column entry for Nobs(Tr.A,Tr.B) indicates that 16 subjects had exactly the same altered endowments in Treatments A and C. Paired Wilcoxon and paired t -tests for these 16 subjects reported in the 2[#] column of the left panel of Table 4 imply rejection of H_o^R in favor of H_a^R at 5% significance, which implies that SM reactions to the prior actions of FMs are a significant determinant of SM behavior in the within-subjects experiment.

We also report censored regressions in Table 5 for amounts returned r by SMs as the dependent variable and amounts sent s and a dummy variable $D^a = 1$ for observations from the moonlighting game as explanatory variables. If there is a significant reaction effect then the parameter estimate for $s \times D^a$ should be significantly positive. As reported in the reactions (left) panel of Table 5, the coefficient estimates for $s \times D^a$ are positive and different from 0 at 1% significance for all three experiments. The estimate is also significant at 1% for the subject-fixed-effect censored regressions for Experiment 2 using within-subject data reported in the 2[#] column. We conclude from the nonparametric and regression tests that reactions to FMs' actions are a significant determinant of SMs' behavior.

4.5 Positive reciprocity is significant

In this section we ask whether subjects' behavior supports the conclusion that there is significant positive reciprocity by comparing data from Treatments A and C. We here examine the behavior of SMs and dictators who have altered endowments larger than the paired subjects' altered endowments (because, in Treatment A, the amount sent by first movers s^a was positive). We test the null Hypothesis H_o^P , that amounts returned are the same in Treatments A and B, against the alternative Hypothesis H_a^P , that amounts returned are larger in Treatment A, as stated in the bottom, middle panel of Table 1.

As with the reactions tests, the reported numbers of observations used in the paired tests are the numbers of distinct values for amounts sent by FMs in Treatment A except for within-subjects tests reported in column 2[#]. The middle panel of Table 4 reports paired t -tests and paired Wilcoxon tests. The null hypothesis is rejected by both tests at 1% or 5% significance with Experiments 1 and 3 data; for Experiment 2 data, both tests in column 2[^] are significant at 10%. The *within-subjects* tests for Experiment 2 data reported in column 2[#] also reject the null hypothesis at 10% significance.

In addition the middle panel of Table 5 reports results from censored regressions of amounts returned r on amounts sent s by FMs and the interaction $s \times D^a$, where $D^a = 1$ for observations from the moonlighting game, using observations for which amounts sent are nonnegative. For all three experiments, estimated coefficients for $s \times D^a$ are positive and significant at 1%. The estimate is also significant at 1% for the subject-fixed-effect censored regressions for Experiment 2 using within-subjects data reported in column 2[#]. We conclude from the nonparametric and regression tests that the reactions by SMs who received positive amounts from FMs in the moonlighting game are positively reciprocal. It is clear that positive reciprocity is economically

Table 5 Censored regressions for reactions

| | Reactions | | | | | | Positive reciprocity (send ≥ 0) | | | Negative reciprocity (send ≤ 0) | | |
|---------------------|---|----------------|-----------|--|-----------|----------------|--|-----------|-----------|--|-----------|-----------|
| | (absolute amounts of money SM returned) | | | (amounts of money SM returned FM sent) | | | (amounts of money SM returned FM sent) | | | (amounts of money SM returned FM took) | | |
| | 1 | 2 [#] | 2 | 3 | 1 | 2 [#] | 2 | 3 | 1 | 2 [#] | 2 | 3 |
| Test for experiment | | | | | | | | | | | | |
| Nobs (L, R) | 60 (8, 0) | 66 (5, 0) | 66 (5, 0) | 67 (4, 1) | 34 (5, 0) | 50 (4, 0) | 50 (4, 0) | 42 (3, 1) | 32 (3, 0) | 18 (1, 0) | 18 (1, 0) | 38 (1, 1) |
| Amount sent | -0.048 | 0.209** | 0.168* | 0.320** | -0.131 | 0.306* | 0.165 | 0.469** | 0.093 | -0.049 | -0.049 | -0.017 |
| (std. err.) | (0.114) | (0.115) | (0.121) | (0.159) | (0.215) | (0.214) | (0.221) | (0.248) | (0.166) | (0.196) | (0.196) | (0.208) |
| Amount sent × | | | | | | | | | | | | |
| Treatment A dummy | 1.099** | 0.428*** | 0.449*** | 0.824*** | 1.437*** | 0.446*** | 0.483*** | 0.923*** | 0.267** | 0.242** | 0.242** | 0.427** |
| (std. err.) | (0.156) | (0.140) | (0.172) | (0.231) | (0.203) | (0.154) | (0.213) | (0.311) | (0.144) | (0.141) | (0.141) | (0.257) |
| Constant | 0.945** | 1.269* | 1.359** | 0.921* | 0.199 | 0.585 | 1.233 | -0.022 | 0.048 | 0.089 | 0.089 | 0.246 |
| (std. err.) | (0.458) | (0.674) | (0.583) | (0.517) | (1.108) | (1.403) | (1.285) | (0.987) | (0.649) | (0.873) | (0.873) | (0.476) |
| Log-likelihood | -144 | -175 | -178 | -182 | -81 | -136 | -140 | -119 | -61 | -31 | -31 | -82 |

***, ** and * denote significance at 1%, 5% and 10%, respectively.

[#]denotes regressions with subject fixed effects.

L and R stand respectively for left-censored and right-censored observations.

significant, as well as statistically significant, from comparison of the mean amounts returned in Treatments A and C reported in Table 4.

4.6 Negative reciprocity has mixed significance

We here examine the behavior of SMs and dictators who have altered endowments smaller than the paired subjects' altered endowments (because, in Treatment A, the FMs took money). We test the null Hypothesis H_o^N , that amounts returned in Treatments A and C are equal, against the alternative Hypothesis H_a^N , that amounts returned are smaller in Treatment A, given in the lower, right panel of Table 1.

Table 4 tests with Experiment 3 data pair Treatments A and C observations of mean amounts returned on the basis of 4 distinct amounts taken by FMs (by applying the same pairing rule used for preceding across-subjects tests). Neither the t -test nor the Wilcoxon test rejects the null hypothesis. Experiments 1 and 2 data are concentrated at -5 amounts sent by FMs; hence Treatments A and C observations cannot be paired by amounts sent (as above). The null hypothesis of no difference in the amounts returned when the FMs took 5 is rejected at 10% significance by a non-paired t -test for both Experiments 1 and 2. The Wilcoxon (actually, Mann-Whitney) test rejects the null hypothesis at 10% significance for Experiment 2 but not Experiment 1. The *within-subjects* tests for Experiment 2 data, reported in column 2[#], also reject the null hypothesis at 10% significance.

The right panel in Table 5 reports estimates of the coefficients for censored regressions using observations for which amounts sent s are nonpositive. The parameter estimates for $s \times D^a$, where $D^a = 1$ for Treatment A observations, are positive and significant at 5% for all three experiments. Table 5 also reports censored regressions with subject fixed effects for the within-subjects Experiment 2 data in column 2[#]; here, the estimated coefficient for the interactive dummy variable is also significant at 5%. We conclude that the test results provide mixed support for the conclusion that the reactions by SMs who received nonpositive amounts from FMs in the moonlighting game are negatively reciprocal. From comparison of the mean amounts returned in Treatments A and C reported in Table 4, it is clear that negative reciprocity has questionable economic significance as well as the mixed statistical significance described above.

5 Comparisons with the Falk et al. experiment

Falk et al. (2000) reports a dyadic design experiment consisting of the moonlighting game and a random first move treatment that provides a motivation control for SMs in the moonlighting game. The payoff protocol is single-blind. FMs in the moonlighting game choose amounts to give to, or take from, anonymously-paired SMs (the "Intentions treatment"). SMs in the moonlighting game make strategy responses: they choose amounts to give to, or take from, paired FMs for every possible choice by a FM. First moves in the control treatment are randomly selected from a probability distribution (the "No-intentions" treatment); the probability distribution used is based on the empirical distribution of choices by FMs in the moonlighting game reported by Abbink et al. (2000). SMs in the control treatment make strategy responses, as they do

in the moonlighting game. Different subjects participated in the two treatments; hence the central test for the significance of reciprocity (“intentions”) is across-subjects, as in our Experiments 1 and 3. In contrast, since strategy responses are used for SMs, the data can be used to make within-subjects comparisons of SM choices made in the moonlighting game conditional on (hypothetical) FM positive (“give”) or negative (“take”) transfers. The central features of the Falk et al. experiment are summarized in the fourth row of Table 2.

5.1 Tests for reciprocity using data from the Falk et al. experiment

Falk et al. reports tests for positive and negative reciprocity using data from the moonlighting game and random move control treatment. They report nonparametric tests that compare the distribution of choices made by subjects in the moonlighting game with choices made by (different) subjects in the control treatment for each hypothetical amount that might be transferred from the first mover. One-sided tests reveal significantly larger absolute values of amounts transferred by SMs in the moonlighting game than in the control treatment. Another test procedure they report uses regression analysis and dummy variables. The dependent variable is the change in FMs’ payoffs caused by SMs’ choices. Explanatory variables are the amount a transferred from a FM to a SM, a dummy variable I for the moonlighting game, and their interaction $a \times I$. The estimated coefficient for $a \times I$ is significant while the other coefficients are not. They conclude that “intentions matter.”

5.2 Comparison of results across studies

The Falk et al. study does not support tests for anticipations, trust, or fear because its dyadic design does not include a control treatment for FMs. Comparisons can be made between the implications of the Falk et al. data and our data for tests of positive and negative reciprocity. Using data from our Experiment 2, with a within-subjects design, we conclude that individual subjects exhibit both positive and negative reciprocity. The aggregate-level tests for reciprocity that are possible with data from the across-subjects designs of our Experiments 1 and 3 and the FFF experiment support somewhat different conclusions. Our tests of aggregate data find significant support for positive reciprocity but mixed support for negative reciprocity. The FFF tests of aggregate data find significant support for both positive and negative reciprocity.

6 Methodological issues

The several experiments now discussed in this paper, together with related experiments reported elsewhere, provide data that can be used to shed light on some methodological issues raised by referees that may be of general interest. We examine the following issues.

6.1 What are the comparative advantages of single blind and double blind protocols in tests for trust and reciprocity?

A concern was that a double-blind protocol might create an experimenter “demand effect” for subjects to be selfish and, furthermore, that this demand effect might be stronger in a dictator game (such as Treatments B and C) than in a strategic game (such as Treatment A), thereby biasing between-treatments comparisons in a triadic experiment *in favor of* detecting trust and reciprocity. This issue can be addressed with data, as follows. If a double-blind protocol “demands” that subjects be selfish, and makes a stronger “demand” on subjects in dictator games than in strategic games, then the triadic design could produce “reciprocity” in Experiment 1, with double-blind payoffs, but not in Experiment 3 with single-blind payoffs. But data from Experiments 1 and 3 support *the same* conclusions about reciprocity in the moonlighting game.

Data do support the conclusion that single-blind and double-blind protocols can lead to different conclusions about reciprocity in a similar game. Cox and Deck (2005), using a triadic design for experiments with the trust game (a simplified form of the investment game), find that subjects exhibit positive reciprocity with a single-blind protocol but not with a double-blind protocol. The reason why Cox and Deck find reciprocity with the single-blind protocol but not with the double-blind protocol is that behavior is different in *the trust game* with a single-blind protocol than with a double-blind protocol; there is no significant difference between behavior in the dictator control treatments with single-blind and double-blind protocols. This is the reverse of the pattern implied by the “experimenter-demand effect” concern about double-blind payoffs. One could interpret this finding as an indicator of an experimenter demand effect from a *single-blind* protocol. Cox and Deck interpret the difference in positive reciprocity in the trust game between double-blind and single-blind protocols as reflecting the difference between a fully internalized social norm for reciprocity and one that is not fully internalized.⁵

Data from our experiments and the FFF experiment tell us more about possible effects of payoff protocols. Table 6 reports tests with data from Experiments 1 and 2 for the joint effects of across-subjects design and double-blind payoffs vs. within-subjects design and single-blind payoffs. The top, left panel of Table 6 reports no significant difference between Treatment B data from Experiments 1 and 2. The bottom, left panel reports tests for differences in Treatment C behavior between experiments using data from our three experiments and data from random-move control treatment (NI) in the FFF experiment. The paired *t*-tests and paired Wilcoxon tests detect no significant differences in behavior from the dictator experiments run with single-blind and double-blind protocols. Therefore the data are uniformly *inconsistent* with the view that double-blind payoff procedures create a demand effect for selfish behavior in dictator games.

We next consider the moonlighting game. The top, right panel of Table 6 reports tests using data on FM decisions. Comparison of Experiments 1 and 2 data show that the joint effects of across-subjects design and double-blind payoffs vs. within-subjects design and single-blind payoffs are insignificant. The simpler comparison for effects

⁵ Cox and Deck (2006) report that female subjects are more responsive than male subjects to a change in the payoff protocol.

Table 6 Effects of procedures on observed behavior

| | First mover behavior | | | |
|------------------------------------|---------------------------|---------------------------|------------------|------------------|
| | Dictator games (Tr:B) | Moonlighting games (Tr:A) | | |
| Nobs (Exp. 1, Exp. 2, Exp. 3) | (27, 33, 0) | (30, 33, 32) | | |
| Means {Exp. 1, Exp. 2, Exp. 3} | {-3.93, -3.73, xxx} | {1.13, 2.73, 0.91} | | |
| Std. dev. (Exp. 1, Exp. 2, Exp. 3) | (3.100, 2.516, xxx) | (5.975, 4.907, 4.306) | | |
| | Exp. 1 vs Exp. 2 | Exp. 1 vs Exp. 2 | Exp. 1 vs Exp. 3 | Exp. 2 vs Exp. 3 |
| Mann-Whitney | -0.321 | -1.160 | -0.214 | 1.961** |
| Smirnov | 0.138 | 0.261 | 0.275 | 0.325** |
| Epps-Singleton | 5.92 | 7.60 | 13.25** | 15.67** |
| Second mover behavior | | | | |
| | Moonlighting games (Tr:A) | | | |
| | Dictator games (Tr:C) | Exp. 1, Exp. 3 | Exp. 2, Exp. 3 | Exp. 3, Exp. FFF |
| Nobs | (7, 7) | (6, 6) | (9, 9) | (9, 9) |
| Means | {0.30, 1.41} | {2.27, 1.38} | {0.19, 0.65} | {2.92, 3.26} |
| (std. dev.) | (0.796, 2.180) | (1.950, 2.387) | (0.843, 0.998) | (5.729, 5.935) |
| <i>t</i> -test (paired) | -1.207 | 0.910 | -1.069 | -0.546 |
| Wilcoxon (paired) | -1.190 | 0.843 | -1.125 | 0.255 |
| | | Exp. 1, Exp. 3 | Exp. 2, Exp. 3 | Exp. 3, Exp. FFF |
| | | (8, 8) | {2.27, 3.11} | (9, 9) |
| | | (3.088, 5.461) | (2.579, 2.349) | {0.98, 1.15} |
| | | -0.639 | -0.380 | -0.380 |
| | | 0.351 | -0.533 | -0.533 |

***, ** and * denote significance at 1%, 5% and 10%, respectively.

Paired tests reported in two bottom rows of the table are run on means of returns observed at given amounts sent.

of changing only the payoff protocol involves data from Experiments 1 and 3. Tests comparing these two experiments have mixed significance; the p -values are 0.00 for Epps-Singleton, 0.10 for Smirnov, and 0.47 for Mann-Whitney. Thus there is only weak support for the conclusion that double-blind payoffs cause first movers to be less generous in the moonlighting game. All tests are significant for the comparison between Experiments 2 and 3; first movers are more generous in Experiment 2 than in Experiment 3. This last comparison confounds effects of the payoff protocol with the effects of subjects having more than one decision because the group X subjects in Experiment 2 knew that there would be subsequent decision rounds in the experiment at the time they made their decisions for the moonlighting game. The possible effects on subjects' behavior of the sequential decisions made by individual subjects in a within-subjects experiment such as Experiment 2 are examined in Section 6.3.

The bottom, right panel of Table 6 presents tests for effects of payoff protocol on SM decisions in the moonlighting game. Neither test is significant for either the comparison between Experiments 2 and 3 or the comparison between Experiment 3 and the FFF experiment.

6.2 Can betrayal aversion invalidate triadic experimental designs in tests for trust?

A first mover who is "betrayal averse" would get disutility from the event of second mover defection in the moonlighting, investment, and other "trust games" in addition to the utility implications of *material costs* from defection (Bohnet and Hong, 2004). A betrayal-averse person would send less to a SM in a trust game than he would send to a paired person in a game against nature in which he faced the same probability distribution of material payoffs (to himself and the other person) as in the trust game.

Suppose that we were to use data from the moonlighting game to construct a game against nature in which FMs faced exactly the same probability distribution of returns that they faced in the moonlighting game. Then betrayal-averse individuals would send more to the paired subject in the game against nature than they would send to SMs in the moonlighting game. An experiment of this type would discriminate between two reasons why an individual might not trust someone else: risk aversion over the distribution of material returns vs. betrayal aversion. Such discrimination, although interesting, has no implication for the validity of the triadic design with dictator control treatments. The FM dictator control along with the moonlighting game provides a test *for the presence* of trusting behavior by asking whether individuals send more to a SM when they might receive a material return (from the SM) than the utility-maximizing amount they choose to send when they know that the other person has no opportunity to return anything. Thus the comparison between Treatment A and Treatment B addresses the question of whether or not the amounts sent in a trust game can be fully explained by altruism or, alternatively, reveal trust in the other person. This comparison is the same, and the correct conclusions about trust are the same, regardless of whether or not individuals would or would not send larger amounts in the game against nature than in the trust game. Betrayal aversion is one of the reasons why individuals may not trust, but its possible empirical validity does not negate tests *for the presence* of trusting behavior. The tests reported in Table 3 support the conclusion that subjects exhibit trust: neither risk aversion nor betrayal aversion nor any other combination of

reasons for not trusting others are sufficiently strong to prevent subjects from sending more money to SMs in the moonlighting game than in a dictator game in which other-regarding preferences are the only reason to send money to another.

6.3 What are the comparative advantages of across-subjects and within-subjects designs in tests for trust, fear, and reciprocity?

Use of both across-subjects and within-subjects designs are well established in experimental economics. The across-subjects vs. within-subjects design issue does not involve a simple dichotomy, as the labels seem to suggest. For example, the Falk et al. experiment with the moonlighting game uses a design that is across-subjects for measuring significance of reciprocity (with their moonlighting game and control treatments) and *both* within-subjects and across-subjects, through use of the strategy method, for measuring the effects of changing endowments.

The principal advantage of a within-subjects design for studying reciprocity is that it focuses directly on individual subjects' reactions to introduction of actions by other subjects, as in our Experiment 2. In contrast, an across-subjects design, such as that of Falk et al. and our Experiments 1 and 3, measures the significance of reciprocity with differences between responses by two distinct samples of subjects from the same population to two treatments that differ only by the presence or absence of actions by the paired subjects. In trust and reciprocity experiments, as in any other type of experiment, the across-subjects design can require a larger sample size than a within-subjects design to identify significant treatment effects because of the different subject characteristics in the two subject samples. A within-subjects design has a different disadvantage: a sequence of choices in fairness games can change behavior. For example, telling subjects there will be another decision task following a dictator game can significantly shift 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 our Experiment 2, a subject knows she has another decision to make after she makes her decision as FM in the moonlighting game. As noted above, the existence of the second decision task following the moonlighting game in Experiment 2 appears to make FMs more trusting and fearful than in Experiment 3 where FMs have no subsequent decision task.

There are significant questions involved in the within-subjects vs. across-subjects choice for experimental design; therefore it is important to learn whether the choice makes a difference for central conclusions about behavior. In our context, one answer is straightforward: tests for trust, fear, and reciprocity using data from the within-subjects Experiment 2 imply the same conclusions as tests using data from across-subjects Experiments 1 and 3.

7 Concluding remarks

Data from experiments reported here and in Falk et al. (2000) have a clear implication for application of the principle of parsimony in modeling social preferences. Unconditional other-regarding preferences models cannot adequately represent behavior in the moonlighting game because of the robust finding that behavior is characterized by

trust and positive reciprocity. Triadic design experiments with the investment game (Cox, 2002, 2004) and the trust game (Cox and Deck, 2005, 2006) also support the significance of trust and positive reciprocity as characteristics of subjects' behavior. Data from experiments reported by Blount (1995), Offerman (2002), McCabe et al. (2003), and Charness (2004) also support the conclusion that reactions to the intentional actions of another are a significant determinant of behavior.

Data from triadic and dyadic design experiments with the moonlighting game support the conclusion that subjects' behavior is characterized by trust and positive reciprocity. The conclusion about trust has been found to be robust to experiments with within-subjects vs. across-subjects designs and single-blind vs. double-blind payoffs. The conclusion about positive reciprocity has been found to be robust to random vs. dictator FM control treatments and strategy responses vs. sequential play as well as within-subjects vs. across-subjects designs and single-blind vs. double-blind payoffs. Support for subjects' negative reciprocity is not as robust to these variations in experimental design and protocol. Subjects' fear of negative reciprocity has not been found to be significant.

Several different types of models represent reciprocal behavior, including the models reported by Levine (1998), Guttman (2000), Rabin (1993), Dufwenberg and Kirchsteiger (2004), Sobel (2005), Cox et al. (in press), and Cox et al. (2006). The empirical-scientific value of these models will be determined by whether they can be successfully applied to data. The parametric model in Cox et al. (in press) has been successfully applied to SM data from Experiment 1, reported here, and data from experiments (by other researchers) with five other types of games. The nonparametric model in Cox et al. (2006) has been successfully applied to SM data from the triadic design experiment with the investment game (Cox, 2004) and data from two other games.

Much work remains to be done in fully capturing the rich behavioral patterns in data for both FMs and SMs in extensive form fairness games such as the ultimatum, investment, trust, and moonlighting games. But two things are clear: reactions to others' intentional actions matter and anticipations of those reactions also matter. Therefore, the principle of parsimony does *not* support the use of unconditional social preferences models for modeling behavior in these games.

Acknowledgments Helpful comments and suggestions were provided by two of the referees. We thank Ingrid Seinen for help in writing the Experiment 1 subject instructions in Dutch and Jos Theleen and Richard Kiser for software programming. We are grateful for research support from the Center for Research in Experimental Economics and Political Decision Making (CREED) at the University of Amsterdam, the Economic Science Laboratory (ESL) at the University of Arizona, and the National Science Foundation (grant numbers SES-9818561 and DUE-0226344).

References

- Abbinck, K., Irlenbusch, B., & Renner, E. (2000). The moonlighting game: An empirical study on reciprocity and retribution. *Journal of Economic Behavior and Organization*, *42*, 265–277.
- Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the rationality of altruism. *Econometrica*, *70*, 737–753.
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142.

- Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63, 131–144.
- Bohnet, I., & Hong, K. (2004). Status and distrust: The relevance of inequality and betrayal aversion. *KSG Faculty Research Working Paper Series*, John F. Kennedy School of Government, RWP04-041.
- Bolton, G., Brandts, J. & Ockenfels, A. (1998). Measuring motivations for the reciprocal responses observed in a simple dilemma game. *Experimental Economics*, 1, 207–219.
- Bolton, G.E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity and competition. *American Economic Review*, 90, 166–193.
- Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics*, 22, 665–688.
- Charness, G., & Rabin, M. (2002). Social preferences: Some simple tests and a new model. *Quarterly Journal of Economics*, 117, 817–869.
- Cox, J. C. (2002). Trust, reciprocity, and other-regarding preferences: Groups vs. individuals and males vs. females. In R. Zwick & A. Rapoport (Eds.), *Advances in experimental business research*, Kluwer Academic Publishers.
- Cox, J. C. (2000) Trust & reciprocity: Implications of game triads and social contexts. *Discussion Paper*, University of Arizona, revised 2003.
- Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46, 260–281.
- Cox, J. C., & Deck, C. A. (2005). On the nature of reciprocal motives. *Economic Inquiry*, 43, 623–635.
- Cox, J. C., & Deck, C. A. (2006). When are women more generous than men? *Economic Inquiry*, 44, 587–598.
- Cox, J. C., Friedman, D., & Gjerstad, S. (in press). A tractable model of reciprocity and fairness. *Games and Economic Behavior*.
- Cox, J. C., Friedman, D., & Sadiraj, V. (2006). Revealed altruism. ExCEN Working Paper Number 2006-09, Georgia State University.
- Cox, J. C., & Sadiraj, V. (2004). Direct tests of models of social preferences and a new model. Unpublished paper, University of Arizona.
- Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47, 268–298.
- Falk, A., Fehr, E., & Fischbacher, U. (2000). Testing theories of fairness: Intentions matter. Working paper No. 65, Institute for Empirical Research in Economics, University of Zurich.
- Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114, 817–868.
- Guttman, J. M. (2000). On the evolutionary stability of preferences for reciprocity. *European Journal of Political Economy*, 16, 31–50.
- Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1, 593–622.
- McCabe, K., Rigdon, M., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, 52, 267–275.
- Offerman, T. (2002). Hurting hurts more than helping helps: The role of the self-serving bias. *European Economic Review*, 46, 1423–1437.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83, 1281–1302.
- Sobel, J. (2005) Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43, 392–436.