Available online at www.sciencedirect.com
SCIENCE DIRECT®
Games and Economic Behavior 53 (2005) 151–169
GAMES and Economic Behavior
www.elsevier.com/locate/geb

Expressed preferences and behavior in experimental games

Gary Charness [a,*], Matthew Rabin [b]

[a] *Department of Economics, 2127 North Hall, University of California, Santa Barbara, CA 93106-9210, USA*
[b] *Department of Economics, 549 Evans Hall, University of California, Berkeley, CA 94720-3880, USA*

Received 26 February 2003

Available online 10 December 2004

Abstract

Participants in experimental games typically can only choose actions, without making comments about other participants' future actions. In sequential two-person games, we allow first movers to express a preference between responder choices. We find that responder behavior differs substantially according to whether first movers express a hope for favorable or unfavorable treatment. Responders largely ignore first movers' expressed preferences for favorable responses, however, when the first movers misbehave. As in earlier experiments without preference expression, subjects assign a high positive weight to another person's payoffs when ahead and misbehavior elicits a strong negative response. Logit regressions estimate the weight placed on another (non-misbehaving) person's payoffs to be positive, even when one is behind. There is suggestive evidence that positive reciprocity is enhanced when a preference for favorable treatment is expressed.

JEL classification: A12; A13; B49; C70; C91; D63

Keywords: Beliefs; Experiment; Expressed preferences; Positive reciprocity; Social preferences

Express yourself.
Madonna

* Corresponding author.
E-mail addresses: charness@econ.ucsb.edu (G. Charness), rabin@econ.berkeley.edu (M. Rabin).
URLs: http://www.econ.ucsb.edu~charness, http://emlab.Berkeley.EDU/users/rabin/index.html.

0899-8256/$ – see front matter
doi:10.1016/j.geb.2004.09.010

1. Introduction

Recent experimental studies have shown that willingness to sacrifice in games is sensitive not only to the choice set available to the player contemplating an action, but also to the behavior of other players that generated that choice set.[1] People are concerned not only with the distribution of material payoffs among players, but also with the process leading to those payoffs. Because a key means of inferring the intentions of another player is to compare the choice made to the set of choices that player could have made but did not, and because sound psychology and common sense tells us that people react to the perceived intentions of others, most researchers interpret the influence of foregone choices on other participants' reactions to mean that participants are influenced by perceived intentions.

Even a comparison of chosen actions to unchosen actions may, however, leave another player's intentions open to interpretation, among other reasons because the intended consequences of a player's action depend on that player's expectations about how others will respond to that action. In most real-world social interactions, an important way people make intentions, preferences, and expectations clear is simple communication, via costless and non-binding messages. People often say why they are doing something and how they hope others will respond.

Most experimental studies of social preferences have, however, not allowed any communication between subjects.[2] This 'no-communication protocol' therefore misses an important element of many of the real-world social and economic interactions that (presumably) experimental games are meant to help us understand. While existing experimental studies have found cheap talk to be very effective in coordinating intended future actions in coordination games, in this paper we explore whether a different form of communication, namely *expressing a preference* between a responder's possible choices, can affect the behavior of the players in simple experimental games. We might well expect expressed preferences by a first mover to affect the responder's beliefs about the first mover's hopes and motivations.

We conduct experiments on a series of simple sequential games, and compare the results among these games and to results from similar games in Charness and Rabin (2002). We compare behavior in binary-choice games in which one player is required to state a preference between two potential responder choices to behavior in the same games when preference expression is not permitted. We use *dictator* games (where one player makes a unilateral allocation) and *response* games (where a first mover has an outside option or can 'enter,' passing the choice to the responder). Foregone options in response games provide inferences regarding the intentions of the first mover, but these inferences may not be rock-solid.

We find that differing expressed preferences lead to significantly different responses when the individual expressing these preferences has not acted unfavorably toward the responder. When the expressing party *has* acted unfavorably, however, expressed preferences for favorable treatment generally fall on deaf ears, and may even be counter-productive. It appears that people are responsive to the explicit hopes of people who have not been un-

[1] For example, see Brandts and Solà (2001), Falk et al. (2003), Charness and Rabin (2002), and Cox (2004).

[2] Exceptions include Brandts and Charness (2003), Charness and Dufwenberg (2002), and (most closely related to our analysis in this paper) Hannan et al. (2002).

kind to them, but are not bothered by disappointing selfish people. This result is in line with models that assume preferences depend in part on beliefs, but is not consistent with purely consequential models.

We use logit regressions on responder behavior to estimate a number of parameters, as in Charness and Rabin (2002). We once again find that the desire to increase someone else's payoff when he or she is behind is a key factor, and that misbehavior by the first mover is a strong and significant influence on responder behavior. We also observe that, on average, people prefer to increase the payoffs of other people who have not misbehaved, even when these others are already receiving more. The coefficient on this parameter is positive in most specifications, and is not significantly negative in any of them; this result is at variance with the presumptions in the Fehr and Schmidt (1999) and Bolton and Ockenfels's (2000) distributional models. We therefore reinforce the perspective that these models not only omit the role of reciprocity in explaining why players hurt others when behind; once reciprocity is accounted for, they have the sign wrong for how the typical subject cares about others' payoffs when behind. Our regressions also confirm a major role for expressed preferences in explaining responder behavior.

Models such as Rabin (1993) and Dufwenberg and Kirchsteiger's (2004) posited both the positive and negative (defined as the difference in responses to another person's favorable or unfavorable action perceived to be intentional vs. a 'neutral' action). Yet while abundant evidence of negative reciprocity has been found, only a handful of the many experimental studies provide any evidence of positive reciprocity.

We test again for positive reciprocity by comparing responses to an identical choice set generated variously by an intentional first-mover choice or by the experimenter, and by introducing a dummy in our regressions. We find that expressing a preference for a favorable response when making a favorable play appears to induce significant positive reciprocity; the magnitude of this effect is nearly as large as the effect from negative reciprocity. This suggests that beliefs about the desires (or expectations) of other players are influential here; perhaps a stronger form of communication would lead to more powerful effects.

In Section 2, we discuss some issues and evidence with respect to beliefs and social preferences in experimental games. The experimental design and results are presented in Section 3, and we analyze the effects of expressed preferences on behavior in Section 4. We present some regression analysis and discussion in Section 5, and conclude in Section 6.

2. Beliefs and social preferences

The ultimatum game (Güth et al., 1982) is the classic experimental example of people sacrificing money to lower the monetary payoff of another person. Many responders react to (disadvantageously) lopsided proposals by rejecting those proposals, so that both people receive nothing rather than the lopsided allocations. In distributional models such as Fehr and Schmidt (1999) and Bolton and Ockenfels's (2000), responders reject unfair proposals because of per se aversion to disparities in relative payoffs. Interpreting these models literally, responders' beliefs about the intentions of proposers are irrelevant.

However, there is abundant experimental evidence that responses are influenced by the options that the first player did not select and, implicitly, the responder's view about the

appropriateness of the choice actually made by the first player. In games similar to ultimatum games, Brandts and Solà (2001) and Falk et al. (2003) find substantial differences in both first-mover and responder behavior according to the foregone first-mover choice. Cox (2004), Cox and Deck (2002), and Cox et al. (2001) also find that foregone options affect behavior in variations on the Berg et al. (1995) "investment game." In many games in Charness and Rabin (2002), we varied the outside option (if any) available to A, while keeping the binary responder choices constant. Here again, there are systematic patterns in first-mover and responder behavior that depend on the payoffs in the outside option. To at least some extent, the awareness of foregone outside options allows a responder to infer intentions.

Perceived intentions may come into play through considerations of *reciprocity*, which a variety of social science disciplines suggest is a basic motivational drive in social interaction. There are a number of studies that demonstrate the importance of negative reciprocity. Kahneman et al. (1986) had As choose between (A, B, C) payoffs of $(5, 5, 0)$ vs. $(6, 0, 6)$ when they knew that Bs had previously chosen an even allocation in an earlier (and independent) dictator game, while Cs had chosen a selfish one. 74% of As chose $(5, 5, 0)$, presumably sacrificing to punish an unfair allocator. Blount (1995) finds that people would generally accept a substantially smaller share of a sum of money when they knew the proposed split was generated by a random mechanism than when generated by the (self-interested) party with whom she would split. Using a gift-exchange design, Charness (2004) observes that an unfavorable allocation intentionally chosen by a self-interested person almost invariably led to a "no-gift" response, but that many participants would contribute something to benefit a "blame-free" person when they received a meager allocation chosen at random.

There are fewer clear demonstrations of the impact of good intentions. McCabe et al. (2003), Cox (2004) and Cox and Deck (2002) find evidence of significant positive reciprocity in a simple "trust" game, and Falk et al. (2001) observe positive reciprocity in a moonlighting game. Offerman (2002) finds some evidence of positive reciprocity, though not to a statistically significant degree; Brandts and Charness (2003) also find modest positive reciprocity in their cheap-talk game. But there are far more studies indicating a lack of positive reciprocity. Bolton et al. (1998, 2000) find evidence against positive reciprocity. In our earlier study (Charness and Rabin, 2002), we found clear evidence of positive reciprocity when helping somebody was costless, but no evidence of positive reciprocity otherwise.

Helpful sacrifice in laboratory games can generally be explained by distributional considerations (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999; Charness and Rabin, 2002). Standard games cited as evidence for positive reciprocity (e.g., the gift-exchange game, the prisoner's dilemma and the trust game) confound altruism, equity, and reciprocity, because good behavior by a first player typically generates a situation where equitable behavior by a second player requires her to help the first. If responses are not conditioned on the initial action, it is unclear why this should be considered to be positive reciprocity instead of simple generosity or altruism. In fact, the models of Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and Charness and Rabin (2002), can all explain the afore-mentioned results without invoking reciprocity.

A number of experimental studies have shown that non-binding pre-play communication ("cheap talk") can be very effective in achieving the Pareto-dominant equilibrium outcome in coordination games. Cooper et al. (1992) find a very high level of coordination with two-way communication and a smaller, but still substantial, level for one-way communication. Charness (2000) finds that pre-play communication is very effective, even when credible coordinating messages could also be interpreted by skeptical others as self-serving.

There are also several experiments in which social preferences are clearly relevant where some form of anonymous communication is permitted. Brandts and Charness (2003) require players in one role to send a statement of intended play in a binary-choice game, where one choice is more favorable to the message receiver. If an unfavorable outcome is reached in the subsequent simultaneous game, a receiver can punish the sender, at a cost. They find that punishment is twice as likely after a deceptive signal than after a truthful one. Charness and Dufwenberg (2002) show that (open-ended) promises improve the likelihood of optimal social choices in a principal-agent environment with hidden action, even when these choices involve personal financial sacrifice. In these studies, it would appear that non-binding communication affects beliefs about the play or expectations of other participants.

Closest to our design are Hannan et al. (2002) who conduct a gift-exchange experiment. In one treatment, each firm submitted a wage offer along with a request for some level of costly effort; all wage/effort combinations were then displayed publicly. Workers who accepted wage offers were not bound by the accompanying requests. Requests appear to increase the level of effort provided, with workers often choosing an effort level intermediate with respect to the minimum allowed and the level requested. We are unaware of any other studies in which a player can, *once her move has been chosen*, express a non-binding preference for a response.[3] Our conjecture was that a first mover's direct statement about her preferences will affect a responder's willingness to make a monetary sacrifice.

Even when a responder should in principle be able to infer a first mover's intentions, we suspect that an expressed favorable preference makes these intentions more salient; it is more difficult to ignore a stated preference than a belief with only implicit support. However, even if we do find that expressed preferences can lead to better social outcomes, we must distinguish whether this result reflects positive reciprocity or some other motivation. For example, in Dufwenberg and Gneezy (2000), player A chooses between an outside option of $(x, 0)$ and letting player B choose $(y, 20 - y)$, where $0 \leqslant y \leqslant 20$.[4] As were then asked to guess the average y chosen by Bs, and, simultaneously, Bs were asked to guess the average guess of As. Subjects were rewarded monetarily for the ex post accuracy of their guesses. While x and y were not correlated, the results do show a strong correlation between y and B's expectation of A's expectation of y; the authors interpret this result to mean that one is averse to "letting down" another person who has acted decently.

[3] Fehr et al. (2001) allow first movers to specify response levels in a sequential prisoner's dilemma game; however, this is not cheap talk. If a first mover invests in verification technology, an unfulfilled request automatically leads to (stochastic) punishment.

[4] There were 5 treatments, where x was variously 4, 7, 10, 13, or 16.

In their context, this suggests that behavior might be different to the extent that first-mover preferences are clarified and expectations brought into sharper focus.[5]

3. Experimental design and results

We conducted 23 sessions, eleven in Berkeley, four in Barcelona, and eight in Santa Barbara. Participants played either four or eight games in a session, and knew that they would be paid according to the outcome generated in one or two of these games, selected at random.[6] In all, 668 people participated in our experiments, each in only one session. Average earnings were around $16 in Berkeley and Santa Barbara, about $11 net of the show-up fee paid, for a session lasting about an hour; average earnings were around $7 in Barcelona, about $4 net of the show-up fee paid for a 40-minute session. Recruiting at Berkeley and Santa Barbara was done primarily through the use of campus e-mail lists; we recruited in Barcelona by posting announcements around campus. Each game was played in two of the sessions. In this way, we hoped to smooth variation over individual sessions, and minimize strong session effects.

After observing and analyzing the results in the first 10 sessions, we designed eight additional games for Session 11 that were chosen to fill in missing conditions whose absence diminished our ability to draw inferences from the original array of games. The Barcelona sessions we designed took place after the Berkeley sessions, and were conducted to complete comparisons with games conducted there and reported by Charness and Rabin (2002). The Santa Barbara sessions replicated the 16 games with expressed preferences that were conducted at Berkeley, but with an option for the first mover to remain silent.[7] We conducted no 'pilot studies'; below we report results on all games we have tested related to the topic of this paper beyond the closely-related games reported by Charness and Rabin (2002).

Prior to each session, packets of instructions and decision sheets were placed face down on desks on both sides of a large room. On entering, a participant could choose any unoccupied desk having a packet. In all games reported here, people on opposite sides of the room were randomly paired, and people were told (truthfully) that they would never be matched twice with the same person in *any* two games. Subjects turned over the top sheet which contained the instructions, which were read aloud to the group. The next sheet was turned over, presenting the first game. Prior to decisions being made in the game, the outcome for every combination of choices was publicly described to the players. Once these combinations were described, a coin was flipped to determine the role for each side

[5] Note that this aversion to letting someone down by going against expressed desires could equally well apply in the Hannan et al.'s (2002) experiment.

[6] Not all games in a session pertained to this experiment. In both Berkeley and Barcelona, we simultaneously conducted experiments on three-player diffusion-of-responsibility games. These games were conducted for what was from the beginning intended as a different project. They provided no confirmation of a type of diffusion of responsibility reported in the psychology literature. Results in these games are available from the first author upon request, and we intend to report them elsewhere.

[7] Participants played a mix of games with and without expressed preferences in the Berkeley and Barcelona sessions.

of the room. In games where two people made decisions, first-mover choices were made and their decision sheets were collected, then second-player choices were made and these sheets were collected. The experimenter received the decision sheets face down and put them, without inspection, in an individual folder. We then proceeded to the next game, repeating this sequence (including a new coin flip). After all games were played, a die was rolled to determine which of the games would be chosen for actual payments and these were calculated. People were paid individually and privately.[8]

A responder (B) was not told prior to making his decision about the decision of the first mover (A). B instead designated a contingent choice, after being told that his decision only affected the outcome if A opted to give the responder the choice, so that he should consider his choice as if A's decision made it relevant for material payoffs.[9] We conducted expressed-preference games in which A had no option to remain silent, as well as games in which A had such an option. In the first case, A was given a choice between $A1$ (outside option), $A2$ with a preference for $B1$, and $A2$ with a preference for $B2$; in the second case, A was given a choice between $A1$ (outside option), $A2$ with a preference for $B1$, and $A2$ with a preference for $B2$, and $A2$ with a choice to not express a preference.[10]

We present our analysis and interpretation of the results and parse the results in a more useful way in Sections 4 and 5. But Tables 1–3 show all the results from this paper. In these and all tables, 100 units of lab money equal $1.00 in the Berkeley and Santa Barbara sessions, and 100 pesetas (worth 57 cents at the time of the experiments, and roughly equivalent in purchasing power to $1 in the US) in the Barcelona sessions.

In these and all other tables in this paper, B's alternatives are connected by a hyphen; the payoff pair to the left (right) of the hyphen is the result when B chooses "Left" ("Right").[11] Where preference expression was mandated, we list B's choice according to whether A requested "help me" vs. "don't help me"; we also provide the results from the "no express" condition.[12]

---

[8] Sample instructions can be found in Charness and Rabin (2004).

[9] We are skeptical that use of this *strategy method* induced dramatically different behavior than would a *direct-response method* in which players make decisions solely in response to other players' decisions. Cason and Mui (1998), Brandts and Charness (2000), and Pezanis-Christou and Sadrieh (2003) find that this difference in elicitation methods does not appear to affect behavior; Güth et al. (2001) find no significant difference (at $p = 0.10$) across elicitation method for any of the 24 pairwise comparisons, either for data from individual games or pooled across games. On the other hand, Brosig et al. (2003) do find more punishment in the "hot" version of their low-cost game, and Blount and Bazerman (1996) find that the form of the strategy method used affects ultimatum game behavior. Nevertheless, we are not aware of any case where a treatment effect found using the strategy method is eliminated when the direct-response method is used. We feel that the issue of when and where it is 'safe' to use the strategy method needs further delineation.

[10] We wrote each game on the board and went over every contingency with the participants. We also emphasized that each B should indicate a choice for every contingency, on the assumption that A has chosen $A2$, and took questions about the procedure.

[11] Throughout the remainder of the paper, we will follow the convention that the A's payoff to the left of the hyphen is greater than the A's payoff to the right of the hyphen, even though the presentation in the laboratory varied.

[12] Note that, in some cases in these tables, we have a different number of observations for B players according to A's expressed preference. This is because, on rare occasions, a B player would fail to fill in his or her choice

Table 1
Games without preference expression

Dictator games	*A*'s play	*B*'s helping *A*
(750,375)–(400,400)	–	13/30 (43%)
(400,400)–(0,800)	–	11/25 (44%)
(750,400)–(375,375)	–	20/26 (77%)
(800,200)–(0,0)	–	11/11 (100%)
(750,400)–(400,400)	–	17/25 (68%)
(2000,400)–(400,400)		9/11 (82%)
(450,350)–(350,450)	–	1/20 (5%)
(*450,350*)–(*350,450*)	–	3/29 (10%)
Response games	***A*'s entering**	***B*'s helping *A***
(800,0); (750,375)–(400,400)	9/27 (33%)	9/27 (33%)
(800,0); (400,400)–(0,800)	11/30 (37%)	9/30 (30%)
(800,0); (750,400)–(375,375)	12/25 (48%)	23/25 (92%)
(450,0); (450,350)–(350,450)	11/27 (41%)	7/27 (26%)
(0,800); (400,400)–(0,800)	11/11 (100%)	3/12 (25%)
(550,550); (750,400)–(400,400)	5/26 (19%)	12/26 (46%)
(550,550); (750,375)–(400,400)	3/25 (12%)	3/25 (12%)
(100,1000); (125,125)–(75,125)	14/26 (54%)	22/26 (85%)
(750,750); (750,400)–(375,375)	1/25 (4%)	16/25 (64%)
(550,550); (750,400)–(375,375)	17/27 (63%)	21/27 (78%)
(400,750); (750,400)–(375,375)	27/30 (90%)	24/30 (80%)
(500,500); (800,200)–(0,0)	16/30 (53%)	28/30 (93%)
(700,300); (800,200)–(0,0)	11/26 (42%)	21/26 (81%)
(100,900); (800,200)–(0,0)	24/25 (96%)	17/25 (68%)
(700,1300); (800,200)–(0,0)	9/25 (36%)	21/25 (84%)
(400,1200); (400,200)–(0,0)	10/25 (40%)	21/25 (84%)

Note: Barcelona games are in italics.

Table 2
Games with preference expression, no silence option

Dictator games	*A*'s preference	*B*'s helping *A*	
	(for *B* to play Left)	*A* hopes Left	*A* hopes Right
(750,375)–(400,400)	18/26 (69%)	8/26 (31%)	3/25 (12%)
(400,400)–(0,800)	25/27 (93%)	16/27 (59%)	10/27 (37%)
(750,400)–(400,400)	16/25 (64%)	12/25 (48%)	6/25 (24%)
(600,600)–(200,700)	19/20 (95%)	14/20 (70%)	5/20 (25%)
(450,350)–(350,450)	15/20 (75%)	5/20 (25%)	6/20 (30%)

(*continued on next page*)

for every contingency, and results from games not chosen for payoffs were not inspected until after the session was completed.

Table 2 (*Continued*)

Response games	*A*'s play & preference			*B*'s helping *A*	
	Out	Enter, L	Enter, R	*A* hopes Left	*A* hopes Right
(800,0); (750,375)–(400,400)	14/25	7/25	4/25	7/24 (29%)	2/25 (8%)
(800,0); (400,400)–(0,800)	20/26	6/26	0/26	16/26 (62%)	4/25 (16%)
(450,0); (450,350)–(350,450)	11/25	11/25	3/25	12/24 (50%)	4/25 (16%)
(100,1000); (125,125)–(75,125)	16/30	13/30	1/30	24/30 (80%)	14/30 (47%)
(550,550); (750,400)–(400,400)	22/30	8/30	0/30	15/28 (54%)	6/26 (23%)
(550,550); (750,375)–(400,400)	20/27	7/27	0/27	4/27 (15%)	1/27 (4%)
(375,1000); (400,400)–(250,350)	4/12	6/12	2/12	9/11 (82%)	11/11 (100%)
(700,1300); (800,200)–(0,0)	7/12	5/12	0/12	11/11 (100%)	11/11 (100%)
(400,1200); (400,200)–(0,0)	21/25	3/25	1/25	21/25 (84%)	19/22 (86%)
(700,200); (600,600)–(200,700)	14/20	6/20	0/20	16/20 (80%)	8/19 (42%)
(750,0); (750,400)–(400,400)	8/20	12/20	0/20	20/20 (100%)	13/20 (65%)
(*750,100*); (*700,500*)–(*300,600*)	23/30	7/30	0/30	16/27 (59%)	7/27 (26%)
(*700,200*); (*600,600*)–(*200,700*)	24/30	6/30	0/30	21/28 (75%)	10/28 (36%)
(*450,0*); (*450,350*)–(*350,450*)	20/29	9/29	0/29	14/25 (56%)	8/26 (31%)

Note: Barcelona games are in italics.

Table 3
Games with preference expression, silence option

Dictator games	*A*'s preference			*B*'s helping *A*		
	Left	Right	None	Hope L	Hope R	No pref.
(750,375)–(400,400)	20/31	9/31	2/31	12/31 (39%)	7/31 (23%)	4/31 (13%)
(400,400)–(0,800)	31/31	0/31	0/31	15/31 (48%)	2/31 (6%)	8/31 (26%)
(750,400)–(400,400)	24/34	5/34	5/34	25/32 (78%)	15/32 (47%)	21/33 (64%)
(600,600)–(200,700)	33/35	1/35	1/35	27/35 (77%)	8/35 (23%)	13/35 (37%)
(450,350)–(350,450)	26/31	1/31	4/31	9/30 (30%)	6/30 (20%)	7/30 (23%)

Response games	*A*'s play & preference				*B*'s helping A		
	Out	In, L	In, No	In, R	Hope L	Hope R	Silence
(800,0); (750,375)–(400,400)	20/35	12/35	0/35	3/35	25/35 (71%)	13/35 (37%)	16/35 (46%)
(800,0); (400,400)–(0,800)	22/31	9/31	0/31	0/31	14/31 (45%)	4/31 (13%)	5/31 (16%)
(450,0); (450,350)–(350,450)	11/30	15/30	4/30	0/30	14/30 (47%)	5/31 (16%)	4/30 (13%)
(100,1000); (125,125)–(75,125)	16/34	15/34	2/34	1/34	27/34 (79%)	17/34 (50%)	24/34 (71%)
(550,550); (750,400)–(400,400)	16/31	15/31	0/31	0/31	16/31 (52%)	10/31 (32%)	13/31 (42%)
(550,550); (750,375)–(400,400)	28/34	6/34	0/34	0/34	6/34 (19%)	2/34 (6%)	1/34 (3%)
(375,1000); (400,400)–(250,350)	9/35	25/35	0/35	1/35	33/35 (94%)	31/35 (89%)	33/35 (94%)
(700,1300); (800,200)–(0,0)	19/31	12/31	0/31	0/31	27/31 (87%)	27/31 (87%)	28/31 (90%)
(400,1200); (400,200)–(0,0)	22/31	8/31	1/31	0/31	30/31 (97%)	30/31 (97%)	30/31 (97%)
(700,200); (600,600)–(200,700)	24/34	10/34	0/34	0/34	24/34 (71%)	11/34 (32%)	14/34 (41%)
(750,0); (750,400)–(400,400)	17/35	17/35	1/35	0/35	30/35 (86%)	13/35 (37%)	24/35 (69%)

While we interpret our results in terms of the main topics of this paper in the next two sections, here we note that comparing the results to identical games in Charness and Rabin (2002) sheds light on the methodological issue of whether people behaved differently with 'role reversal'—where subjects all played both roles of each game (against different partners)—in Charness and Rabin (2002) and without role reversal here. Charness and Rabin (2004) reports comparisons from identical games without preference expression in the two studies, with 11 comparisons for *B* play and 7 comparisons for *A* play. There does not appear to be any real pattern of different behavior. None of the 18 comparisons has a difference significant at the 5% level (two-tailed test). Overall, we see no evidence that people make different choices when role reversal is used in the experimental design.

4. Effects of expressed preferences

A clear overall pattern in our data is that responder behavior is quite sensitive to the first player's expressed preference. This is particularly true when *A*'s decision to give *B* a choice is favorable to *B*. On the other hand, expressing a preference for help following selfish or hurtful behavior is ineffective (or even slightly counterproductive) compared to the silent game (no expressed preferences).

The simplest test to whether responder play is sensitive to the expressed preference per se is whether *A*'s expression matters in a dictator game, where *A* has had no choice of action. Table 4 presents this evidence.

The proportion of *B* players who maximize *A*'s payoff is typically considerably higher following "Help me" than following "Don't help me." The differences in *B* play according to whether preference expression was are significant in four of the five individual games at $p = 0.05$ or better, both with and without the silence option; if we aggregate the data by

Table 4
Dictator games and preferences

Game	*B*'s helping *A* (by *A* preference) (in %)							
	No express		Help me		Don't help me		Aggreg. pref.[a]	
	Can't express	Silence chosen	Silence option	No option	Silence option	No option	Silence option	No option
(750,375)–(400,400)	43	13	39	31	23	12	32	25
(400,400)–(0,800)	44	26	48	59	6	37	48	58
(750,400)–(400,400)	68	64	78	48	47	24	71	39
(600,600)–(200,700)	73*	37	77	70	23	25	74	68
(450,350)–(350,450)	5	23	30	25	20	30	29	26
Aggregated total	*48*	*33*	*55*	*47*	*24*	*26*	*51*	*43*

[a] Here, and in later tables, the *aggregated preference outcomes* were calculated as follows: Multiply the proportion of those *A*s expressing preferences for Help by the proportion of *B*s then choosing Help, do the same for those *A*s expressing preferences for Don't Help, and for *A*s choosing silence.

* Entries refer to data from Charness and Rabin (2002).

Table 5
Favorable A play and preferences

Game	B's sacrificing to help A (by A preference) (in %)							
	No express		Help me		Don't help me		Aggreg. pref.	
	Can't express	Silence chosen	Silence option	No option	Silence option	No option	Silence option	No option
(800,0); (750,375)–(400,400)	33	46	71	29	37	8	66	21
(800,0); (400,400)–(0,800)	30	16	45	62	13	16	45	62
(450,0); (450,350)–(350,450)	26	13	47	50	16	16	40	43
(700,200); (600,600)–(200,700)	78*	41	71	80	32	42	71	80
Aggregated total	*43*	*30*	*59*	*54*	*25*	*19*	*55*	*51*

* Entries refer to data from Charness and Rabin (2002).

column, the difference is significant at $p = 0.00$.[13,14] Thus, B choices are generally sensitive to A's expressed preferences, even though these are unaccompanied by any action.

The rate of B's helping behavior is substantially higher without preference expression than when A says "Don't help me," by 48 to 33% ($p = 0.01$, two-tailed test). For A to do as well with expressed preferences as without them, he or she must express a preference for help. Doing so leads to a slightly better aggregated helping rate with the silence option and a slightly worse rate without the silence option. The difference between aggregated rates with the silence option and without this option is not statistically significant.

Turning to response games, we analyze the games by category. We first examine the case where A has made a favorable play—those games where (no matter how B responds) A's choice to enter raises in B's payoff, while lowering A's own payoff (Table 5).[15,16]

It is easy to see that when A has made a favorable choice, a responder is much more likely to help when A expresses a preference for help than when A expresses a preference for no help or chooses silence. This is true for each of the 12 comparisons (eight for the

[13] Throughout this and subsequent sections, the p-value is approximated to two decimal places (so that the statement "$p = 0.00$" really means that $p < 0.005$) and is calculated from the test of the equality of proportions, using the normal approximation to the binomial distribution (see Glasnapp and Poggio, 1985), and assuming that each binary choice is independent. When we have an ex ante directional hypothesis, we use a one-tailed test. Where there is no directional hypothesis, we use a two-tailed test.

[14] It is not clear that aggregating across different games is completely appropriate. However, we do wish to note that our sessions were designed so that each participant was faced with a certain "type" of game at most once. No more than four of our experimental games were played in any session. Typically, one of these games involved favorable entry by A, one game involved unfavorable entry by A, and one game was a dictator game. B never faced the same binary choice more than once in a session. Thus we argue that, for statistical purposes, each observation in our aggregated comparisons within Table 3 and within other tables organized by categories is largely independent.

[15] In one of the games, only one of B's two responses lowers A's payoff, while the other leaves A's payoff the same as if A had not entered.

[16] In this table and others, we occasionally use results from the identical games in our earlier study. These games were played under the same conditions and in the same location as in the current study; the only difference is the role-reversal issue that is found to not make a behavioral difference. We include these games to permit a larger number of comparisons, as not all pertinent games were rerun in our current study.

silence-option case and four in the no-option case), each of which is individually significant at $p = 0.05$ (one-tailed tests). The difference in B's (aggregated) behavior is highly significant ($p = 0.00$ in each case) for help vs. no help and help vs. voluntary silence (in the silence-option condition) and for help vs. no help (in the no-option condition). Once again, we see that, in the aggregate, A does worse by choosing silence than when silence is mandated (30 vs. 43%), and this difference is significant at $p = 0.03$ (two-tailed test). There is little difference in aggregated B behavior between the silence-option and no-option cases.

When A has made a favorable choice, an expressed preference for help substantially improves the likelihood of a favorable response compared to when preference expression is not permitted. In the aggregate, the difference in favor of the silence option (59 vs. 43%) is quite statistically significant ($Z = 2.53$, $p = 0.00$, one-tailed test), while the difference without the silence option (54 vs. 43%) is only marginally so ($Z = 1.61$, $p = 0.05$, one-tailed test).

There are also six games where A's entry is unfavorable to B. We break these down into two subcategories, depending on whether a favorable (to A) response is costly for B (Table 6).

In the first three games, it costs B little or nothing to help A. Here again, B is more likely to help when A requests help than when A requests no help or chooses silence; this is true for each of the nine comparisons. We had no directional hypothesis and are in fact a bit surprised by this. The difference in B's (aggregated) behavior is highly significant ($p = 0.00$ in each case) for help vs. no help in both the silence-option and no-option conditions, while it is marginally significant ($p = 0.11$, two-tailed test) for help vs. voluntary silence. Once again, there is little difference in aggregated B behavior between the silence-option and no-option cases, and A does worse by choosing silence than with mandated silence, although this difference is not significant.

Table 6
Unfavorable A play and preferences

Games where B can punish A at no cost	B's helping A (by A preference) (in %)							
	No express		Help me		Don't help me		Aggreg. pref.	
	Can't express	Silence chosen	Silence option	No option	Silence option	No option	Silence option	No option
(550,550); (750,375)–(400,400)	12	3	19	15	6	4	19	15
(100,1000); (125,125)–(75,125)	85	71	79	80	50	47	76	78
(550,550); (750,400)–(400,400)	46	42	52	54	32	23	52	54
Aggregated total	*48*	*38*	*49*	*51*	*29*	*25*	*47*	*49*
Games where B must pay to punish A	B's hurting A (by A preference) (in %)							
	No express		Don't hurt me		Hurt me		Aggreg. pref.	
	Can't express	Silence chosen	Silence option	No option	Silence option	No option	Silence option	No option
(700,1300); (800,200)–(0,0)	16	10	13	0	13	0	13	0
(400,1200); (400,200)–(0,0)	16	3	3	16	3	14	3	15
(375,1000); (400,400)–(250,350)	12*	6	6	18	11	0	7	14
Aggregated total	*14*	*6*	*7*	*13*	*9*	*7*	*8*	*10*

* Entries refer to data from Charness and Rabin (2002).

Table 7
Positive reciprocity without expressed preferences

Games without expressed preferences	*B*'s helping *A* (in %) when:		*Z*
	B is Dictator	*A* has a play	
(800,0); (400,400)–(0,800)	44	30	−1.07
(800,0); (750,375)–(400,400)	43	33	−0.77
(450,0); (450,350)–(350,450)	5	26	1.89
(750,0); (750,400)–(400,400)	68	94*	2.74
(*450,0*); (*450,350*)–(*350,450*)	10	6*	−0.72
Aggregated total	*32*	*39*	**1.20**

* Entries refer to data from Charness and Rabin (2002).

There is no real pattern in the three games where *B* can choose a costly punishment, as punishment rates are similarly low in all cases. Responders do not 'respect' *A*'s preference when he or she has acted unfavorably by entering.

We can also consider the effect of expressed preferences for unfavorable treatment. These results can be extracted from earlier tables, so we only present the detail in Charness and Rabin (2004). When *A* acts favorably and asks for an unfavorable response, *B* is significantly less likely to sacrifice to help *A* in both the silence-option and no-option conditions. When the responder can reduce *A*'s payoff at no cost to herself, she is significantly more likely to do so when *A* expresses this preference. However, when reducing *A*'s payoff is costly, this preference actually may reduce the likelihood of punishment, although these differences are not statistically significant.[17]

We next examine whether preference expression helps to induce positive reciprocity. To test for positive reciprocity, we compare B behavior after a favorable *A* play to *B* behavior in the dictator version of the binary choice, holding constant *whether or not preference expression was a feature*. We consider the two cases in games where entry by *A* is favorable to *B*, first examining the results without expressed preferences (Table 7).

The aggregated data show a slight overall tendency toward positive reciprocity. But this is not statistically significant and, moreover, the sign would be reversed if we exclude the results from the game where *B* can help *A* without sacrificing any money. We see that *B* is actually slightly *less* likely to help *A* after a favorable play in three of the five games. Thus, we cannot conclude that we observe positive reciprocity when expressed preferences are not permitted.

Turning to behavior with expressed preferences, we see a rather different picture (see Table 8).

Overall, there appears to be a significant ($p = 0.04$ and 0.01, one-tailed tests) tendency for positive reciprocity per se to be triggered by a stated preference for favorable treatment; this applies to both preference-expression conditions. Once again, the biggest differences are found when *B*'s help costs nothing or very little. Expressed preferences do appear

[17] Expressing a preference has no significant effect on the likelihood of *A* choosing to enter either when entry is favorable to *B* or entry is unfavorable to *B*. Note that, in general, *A* does not treat preference expression capriciously, as it is highly unusual (7 of 71 cases) for *A* to express a preference for less money when making an entry choice favorable to *B*.

Table 8
Positive reciprocity with expressed preferences

Entering is a favorable play and A prefers a favorable response	B's helping A (in %)				Z_S	Z_{NO}
	Silence option		No option			
	Dict.	Pref.	Dict.	Pref.		
(800,0); (400,400)–(0,800)	48	45	59	62	−0.25	0.17
(800,0); (750,375)–(400,400)	39	71	31	29	2.67	−0.12
(450,0); (450,350)–(350,450)	30	47	25	50	1.33	1.70
(700,200); (600,600)–(200,700)	77	71	70	80	−0.62	0.73
(750,0); (750,400)–(400,400)	78	86	48	100	0.81	2.79
Aggregated total	*55*	*65*	*47*	*62*	*1.75*	*2.31*

to make good intentions salient to some degree, although the bulk of the improvement in B's behavior would appear to stem from an unwillingness to go against the hopes (or expectations) of someone who has not misbehaved.[18]

5. Regression analysis

We turn now to an approach to summarizing our data that assumes that all subjects share a fixed set of preferences, and that observed behavior corresponds to individuals implementing those preferences with error. The likelihood of error is assumed to be a decreasing function of the utility cost of an error. We adapt the simple conceptual model of social preferences in two-person games presented in Charness and Rabin (2002), adding parameters for positive reciprocity and expressed preferences, as well as an interaction term. Letting π_A and π_B be player A's and B's money payoffs, consider the following formulation of player B's preferences:

$$U_B(\pi_A, \pi_B) \equiv (\rho \cdot r + \sigma \cdot s + \theta \cdot q + \tau \cdot p + \upsilon \cdot n + \omega \cdot m + \psi \cdot p \cdot n) \cdot \pi_A + (1 - \rho \cdot r - \sigma \cdot s - \theta \cdot q\tau \cdot p - \upsilon \cdot n - \omega \cdot m - \psi \cdot p \cdot n) \cdot \pi_B,$$

where

$r = 1$ if $\pi_B > \pi_A$, and 0 otherwise;
$s = 1$ if $\pi_B < \pi_A$, and 0 otherwise;
$q = 1$ if A has misbehaved, and 0 otherwise;
$p = 1$ if A has helped B by entering, and 0 otherwise;
$n = 1$ if A has hoped for B to be nice, and 0 otherwise;
$m = 1$ if A has hoped for B to be mean, and 0 otherwise.

[18] One issue that interested us is whether preference expression produces better social outcomes. For response games in which A has acted favorably by entering, we see that preference expression improves aggregate social outcomes (expected efficiency and expected minimum payoffs) are improved by about 20–30%. However, preference expression does not improve social outcomes in either punishment games or dictator games. See Charness and Rabin (2004) for more detail.

This formulation says that B's utility is a weighted sum of her own material payoff and A's payoff, where the weight B places on A's payoff may depend on whether A is getting a higher or lower payoff than B and on whether A has behaved unfairly.[19] The parameters ρ, σ, θ, and τ capture various aspects of social preferences. The parameters ρ and σ allow for a range of different 'distributional preferences,' relying solely on the outcomes and not on any notion of reciprocity; in the difference-aversion models, ρ is positive and σ is negative, while both are positive in the distributional form of the Charness and Rabin (2002) quasi-maximin model. The parameters τ and θ provide a means for modeling positive and negative reciprocity, respectively.

We estimate the population means for our parameters by performing maximum-likelihood estimation on our binary-response data. In this approach, the logit regression

$$P(\text{action1}) = \frac{e^{\gamma \cdot u(\text{action1})}}{e^{\gamma \cdot u(\text{action1})} + e^{\gamma \cdot u(\text{action2})}}$$

determines the values that best match predicted probabilities of play with the observed behavior.[20]

We first consider a logit regression on only the data from dictator games, thereby avoiding possible confounds from reciprocity considerations. Three of the coefficients are found to be significant: ρ is estimated to be 0.344 ($t = 9.28$), υ is estimated to be 0.125 ($t = 2.63$), and ω is estimated to be -0.240 ($t = -3.93$); σ is 0.003, not significantly different from zero. Thus, this regression provides no support for the Fehr and Schmidt (1999) and Bolton and Ockenfels's (2000) models, as sacrifice in these models is driven by negative values of σ.

We next consider regressions on the full data sample.[21] Table 9 reports regression results under a spectrum of different restriction assumptions.

[19] Another way of writing this utility function that some readers might find more intuitive is to break it down into two cases:

$$\begin{aligned}
\text{for } \pi_B \geqslant \pi_A, \quad U_B(\pi_A, \pi_B) &\equiv (1 - \rho - \theta q - \tau p - \upsilon n - \omega m - \psi p \cdot n)\pi_B \\
&\quad + (\rho + \theta q + \tau p + \upsilon n + \omega m + \psi p \cdot n)\pi_A; \\
\text{for } \pi_B \leqslant \pi_A, \quad U_B(\pi_A, \pi_B) &\equiv (1 - \sigma - \theta q - \tau p - \upsilon n - \omega m - \psi p \cdot n)\pi_B \\
&\quad + (\sigma + \theta q + \tau p + \upsilon n + \omega m + \psi p \cdot n)\pi_A.
\end{aligned}$$

[20] The precision parameter γ reflects sensitivity to differences in utility, where the higher the value of γ, the sharper the predictions. When γ is 0, the probability of either action must be 50 percent; when γ is arbitrarily large, the probability of the action yielding the highest utility approaches 1. This approach assumes that all subjects share a fixed set of preferences, and that observed behavior corresponds to individuals implementing those preferences with error. The likelihood of error is assumed to be a decreasing function of the utility cost of an error.

[21] We do not show the specifications that include dummies to test for differences between the silence-option games and the no-option games. By and large, there are no significant differences in the estimated coefficients for our parameters, although the coefficient for ρ is slightly lower and the coefficient for σ is slightly higher with the silence option, both by less than 0.1 compared to the baseline with no silence option.

Table 9
Regression estimates for B behavior ($N = 3053$)

Model	Restrictions	ρ	σ	θ	τ	υ	ω	ψ	γ	LL
Self-interest	$\rho = \sigma = \theta = \tau = \upsilon = \omega = \psi = 0$	–	–	–		–	–	–	0.004 (17.5)	−1979.0
ρ only	$\sigma = \theta = \tau = \upsilon = \omega = \psi = 0$	0.344 (29.0)	–	–		–	–	–	0.010 (18.8)	−1872.8
ρ, σ only	$\theta = \tau = \upsilon = \omega = \psi = 0$	0.345 (29.6)	−0.024 (−1.82)	–		–	–	–	0.010 (18.8)	−1871.2
ρ, σ, θ	$\tau = \upsilon = \omega = \psi = 0$	0.351 (34.1)	0.028 (1.70)	−0.127 (−5.02)		–	–	–	0.011 (19.3)	−1858.9
$\rho, \sigma, \theta, \tau$	$\upsilon = \omega = \psi = 0$	0.346 (25.8)	0.024 (1.32)	−0.123 (−4.65)	0.010 (0.59)	–	–	–	0.011 (19.3)	−1858.8
$\rho, \sigma, \theta, \tau, \upsilon$	$\psi = 0$	0.348 (21.6)	0.032 (1.69)	−0.123 (−4.93)	0.013 (0.76)	0.110 (6.05)	−0.150 (−7.57)	–	0.012 (19.9)	−1776.5
$\rho, \sigma, \theta, \tau, \upsilon, \omega$	None	0.361 (21.8)	0.042 (2.17)	−0.125 (−4.97)	−0.027 (−1.30)	0.070 (3.18)	−0.145 (−7.24)	0.100 (3.02)	0.012 (19.7)	−1772.3

Notes: t-statistics are in parentheses; γ is the precision parameter, and LL is the log-likelihood function.

In all the regressions, coefficient ρ is estimated to be around 0.35 and is always highly significant. Coefficient σ is always small, ranging from -0.024 to 0.042; it is significantly *positive* in the bottom regression, the opposite of the sign predicted by the difference-aversion models. We see a highly significant negative coefficient on the dummy for Hope Mean (υ), and a significantly positive coefficient for the Hope Nice dummy (ω). The coefficient for negative reciprocity is about -0.125 in all specifications, and is highly significant. The coefficient for positive reciprocity is slightly positive in some specifications, but never significant. It is interesting to see that favorable A entry combined with an expressed preference for favorable treatment leads to a substantial and significant effect, seen in the coefficient for ψ.[22]

6. Conclusion

Behavior by responders in our simple games is quite sensitive to the preference expressed by first movers. The greatest effects observed result from the clarification of intentions behind a favorable play. Responders are more likely to help first movers after a favorable play when there has been a preference expressed for help than when no such

[22] The most rigorous test for determining the significance of each parameter is to restrict that parameter to a value of zero in the otherwise complete and unrestricted regression. For B behavior, likelihood-ratio tests on parameter restrictions confirm significant effects for ρ, ω, θ, υ, ψ, and σ, but not for τ. Detailed test results can be found in Charness and Rabin (2004).

preference can be indicated. Expressed preferences by *A* affect *B*'s behavior even when *A* has made no play, so to some extent responder behavior reflects the expressed hope (and perhaps the perceived expectations) of the first mover.

There is no role for these effects in either standard theory or in current prominent models of social motivation. It is obvious that the consequentialist models of pure distribution ignore expressed preferences, as nothing occurring before the responder's choice between outcomes is relevant to the choice. The Falk and Fischbacher (in press), Charness and Rabin (2002), and Cok and Friedman's (2002) models combine reciprocity preferences with distributional preferences, but in these models a player's intentions are discerned by comparing the action chosen with the feasible outcome space, without regard for a player's stated preferences.

There are social benefits to expressed preferences, when a favorable action is accompanied by a hope for favorable treatment. Using two different measures of social welfare, we find that optimal outcomes are achieved substantially more frequently with preference expression, whether or not a silence option is included. On the other hand, social welfare in punishment games and in dictator games is not substantially affected by preference expression.

In the aggregate, we do not observe significant positive reciprocity per se for a favorable first-mover play without preference expression. However, we do see a form of positive reciprocity when a favorable play is accompanied by a preference for a favorable response. We feel it is quite possible that a more personal form of communication might well lead to stronger effects. Nevertheless, the primary source of the benefits found seem to stem from the preference expression itself, rather than from any increased salience of the favorable move made by player *A*. Perhaps some of the effect stems from a reluctance to disappoint the first mover, as seems to be the case in the evidence from expressed preferences in dictator games.

Communication is an important element in real-world social and economic interactions. Experimental designs disallowing communication are thereby neglecting a key issue. While experimenters should surely maintain careful control over the communication protocol, gathering more data from high-communication experiments seems potentially very fruitful.

Acknowledgments

We thank Manuel Fernandez, Brit Grosskopf, and Ellen Quarles for help with the sessions, and Ellen Quarles for her help with the data. We thank Jim Cox, Martin Dufwenberg, Lise Vesterlund, and three anonymous reviewers for their helpful comments and suggestions. Charness thanks the MacArthur Foundation for support and experimental funding, and Rabin thanks the Russell Sage, MacArthur, and National Science (Award 9709485) Foundations. This research was begun while Charness was affiliated with Universitat Pompeu Fabra.

References

Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games Econ. Behav. 10, 122–142.
Blount, S., 1995. When social outcomes aren't fair: the effect of causal attributions on preferences. Organ. Behav. Human Dec. Process. 63, 131–144.
Blount, S., Bazerman, M., 1996. The inconsistent evaluation of absolute versus comparative payoffs in labor supply and bargaining. J. Econ. Behav. Organ. 30, 227–240.
Bolton, G., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. Amer. Econ. Rev. 90, 166–193.
Bolton, G., Brandts, J., Katok, E., 2000. How strategy sensitive are contributions? A test of six hypotheses in a two-person dilemma game. Econ. Theory 15, 367–387.
Bolton, G., Brandts, J., Ockenfels, A., 1998. Measuring motivations for the reciprocal responses observed in a simple dilemma game. Exper. Econ. 1, 207–219.
Brandts, J., Charness, G., 2000. Hot vs. cold: Sequential responses in simple experimental games. Exper. Econ. 2, 227–238.
Brandts, J., Charness, G., 2003. Truth or consequences: an experiment. Manage. Sci. 49, 116–130.
Brandts, J., Solà, C., 2001. Reference points and negative reciprocity in simple sequential games. Games Econ. Behav. 36, 138–157.
Brosig, J., Weimann, J., Yang, C.-L., 2003. The hot versus cold effect in a simple bargaining experiment. Exper. Econ. 6, 75–90.
Cason, T., Mui, V., 1998. Social influence in the sequential dictator game. J. Math. Psych. 42, 248–265.
Charness, G., 2000. Self-serving cheap talk: A test of Aumann's conjecture. Games Econ. Behav. 33, 177–194.
Charness, G., 2004. Attribution and reciprocity in an experimental labor market. J. Lab. Econ. 22, 665–688.
Charness, G., Dufwenberg, M., 2002. Promises and partnership. Mimeo.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quar. J. Econ. 117, 817–869.
Charness, G., Rabin, M., 2004. Expressed preferences and behavior in experimental games. Mimeo.
Cooper, R., DeJong, D., Forsythe, R., Ross, T., 1992. Communication in coordination games. Quart. J. Econ. 53, 739–771.
Cox, J., 2004. How to identify trust and reciprocity: Implications of game triads and social contexts. Games Econ. Behav. 46, 260–281.
Cox, J., Deck, C., 2002. On the nature of reciprocal motives. Mimeo.
Cox, J., Friedman, D., 2002. A tractable model of reciprocity and fairness. Mimeo.
Cox, J., Sadiraj, K., Sadiraj, V., 2001. Trust, fear, reciprocity, and altruism. Mimeo.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. 30, 163–182.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Falk, A., Fischbacher, U., in press. A theory of reciprocity. Games Econ. Behav..
Falk, A., Fehr, E., Fischbacher, U., 2001. Driving forces of informal sanctions. Mimeo.
Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Econ. Inquiry 41, 20–26.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 769–816. according to cover; 817-868 in truth.
Fehr, E., Klein, A, Schmidt, K., 2001. Fairness, incentives and contractual incompleteness. Mimeo.
Glasnapp, D., Poggio, J., 1985. Essentials of Statistical Analysis for the Behavioral Sciences. Merrill, Columbus, OH.
Güth, W., Huck, S., Müller, W., 2001. The relevance of equal splits in ultimatum games. Games Econ. Behav. 37, 161–169.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum game bargaining. J. Econ. Behav. Organ. 3, 367–388.
Hannan, L., Kagel, J., Moser, D., 2002. Partial gift exchange in an experimental labor market: Impact of subject population differences, productivity differences, and effort requests on behavior. J. Lab. Econ. 20, 923–951.
Kahneman, D., Knetsch, J., Thaler, R., 1986. Fairness and the assumptions of economics. J. Bus. 59, S285–S300.
McCabe, K., Rigdon, M., Smith, V., 2003. Positive reciprocity and intentions in trust games. J. Econ. Behav. Organ. 52, 267–275.
Offerman, T., 2002. Hurting hurts more than helping helps. Europ. Econ. Rev. 46, 1423–1437.

Pezanis-Christou, P., Sadrieh, A., 2003. Elicited bid functions in (a)symmetric first-price auctions. Mimeo.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Amer. Econ. Rev. 83, 1281–1302.